Indian Psychological Review
1967, Vol. 4, No. 1, 1—17.

The Application of Statistical Sampling Theory to Educational Evaluation*

Rhea S. Das
Indian Statistical Institute, Calcutta.

Populations of students and universes of questions have been considered separately and together in terms of statistical sampling theory and applications to educational evaluation. Sampling a population of students permits a rapid, economical, and accurate estimation of performance, as is required to set up local or national norms and to carry out detailed studies of examination performance. Sampling universes of questions permits objective and rapid setting of examination questions and a statistical basis for estimating a single student's true knowledge. Being based on sampling theory, confidence intervals can be obtained and the probability of classificatory decisions can be estimated for a single student. In this way, sampling theory provides information about individual performance which is unobtainable by conventional psychometric theory. These theoretical results are illustrated by original empirical studies.

Simultaneous sampling of the student population and question universe calls for conceptualization of population and sample matrices. The general element of the population matrix is a random variable; sample estimates of the random variable can be used to compare different student populations or question universes and can serve as indices of changes occurring with increasing academic level and intellectual development.

Not only can educational evaluation be made more objective, more rapid and more accurate through the application of statistical sampling theory, but also the theory and practice of psychometric and educational evaluation can be advanced.

Introduction

As an introduction to the application of statistical sampling theory to educational evaluation, some essential concepts will be briefly considered. Initially three concepts will be mentioned: unit, population, and sample. The term *unit* refers to an element which can be observed and is regarded as individual and indivisible for the purpose of observation. The aggregate of units, that is, the finite or infinite collection of all units, comprises the *population*. A part of the population, or a fraction of the units which comprise the population, deliberately selected according to a specified procedure, forms the *sample*. When the sample has been selected by a procedure based on the theory of probability, it is known as a probability sample, and the selection procedure is referred to as probability sampling. Statistical sampling theory deals with probability sampling and inferences about the characteristics of populations from observations of probability samples, and provides a measure of the precision of the estimates made for the population.

The objective of educational evaluation may be expressed as the accurate assessment of students' academic proficiency. Educational evaluation is generally carried out by requiring students to answer questions and then assessing or marking their answers. The questions may be university or school examination papers or they may be objective achievement or attainment tests. The answers, correspondingly, may be essay-or-limited answer type. Assessment of answers to any given question

* Presented at the Seminar on 'Psychology as a Theoretical and Applied Discipline and its Relationship to allied Natural Scince Disciplines' sponsored by the University Grants Commission at Gujarat University, Ahmedabad, 21—26 March 1966. Published with the permission of Dr. P. H. Prabhu, Director of the Seminar.

may be quantitative (e. g., "marks") or qualitative (e. g., "right" or "wrong"). How can statistical sampling theory be applied to educational evaluation? Two applications, which can be distinguished in terms of the unit observed, will be considered in the following discussion. The units are student and question, and the populations to be sampled are respectively those of students and questions. The terminology and definitions for some common concepts will be presented for the two applications, after which the procedures and potential utility of the applications will be discussed.

Terminology

In this section concepts used in sampling and estimation will be introduced and sampling terminology will be presented for simple random sampling without replacement. Estimation terminology will be given separately for the two populations in subsequent sections.

The concepts of unit, population and sample have already been introduced. The two populations to be considered in educational evaluation are distinguished in terms of unit in column (2) of Table 1. The descriptive names for populations and samples are shown in columns (3) and (7) of Table 1. The appropriate notation for the unit in the population and the sample is given in Table 1 under the headings: general element, number, and subscript. It is necessary to denote the general element, or any unit, of the population (column (4)) and of the sample (column (8)). The number of such elements or units in the population (column (5)) and in the sample (column (9)) must also be indicated. Then, the subscript of the general element (columns (6) and (10)) specifies which element or unit is being considered. Single subscripts are introduced in Table 1.

Table 1
Sampling Notation for Students and Questions

Sl. No.	Description of Unit	Population: Name	Population: General Element	Population: Number	Population: Subscript
(1)	(2)	(3)	(4)	(5)	(6)
1	A Student	Population of Students	E_h	N	$h=1, 2, \ldots, N$
2	A Question	Uuiverse of Questions	U_j	M	$j=1, 2, \ldots, M$

Sample: Name	Sample: General Element	Sample: Number	Sample: Subscript
(7)	(8)	(9)	(10)
sample of students	e_h	n	$h=1, 2, \ldots, n$
sample of questions	u_j	m	$j=1, 2, \ldots, m$

Reference will be made to two sampling procedures in the subsequent discussion, simple random sampling without replacement and stratified random sampling without replacement. Simple random sampling without replacement refers to a probability scheme in which the units are selected with equal probability but, as the selected units are not replaced, it is not possible for the same unit to be selected more than once. With the help of two concepts, sampling frame and stratum, simple random sampling and stratified sampling can be distinguished. The list of all units belonging of to the population, suitably identified, comprises the sampling frame. Sampling the population involves selection of units as listed in the sampling frame according to the probability scheme upon which the sampling procedure is based. The sampling frame may be a unitary or undivided listing of units, allowing simple random sampling. If the units in the population are placed in categories, and listed by categories in the sampling frame, then stratified smapling is possible. The categories of the sampling frame are strata, each stratum defining a homogeneous group of units. Stratified sampling consists of allocating units of the population into strata and selecting samples independently and at random from each stratum. It might be said that whereas all units are sampled in one operation in simple random sampling, they are sampled in several operations, one for each stratum, in stratified sampling. Alternatively, it might be stated that stratified sampling consists of a series of simle random samplings, one for each stratum. Simple random sampling and stratified sampling can be carried out with or without replacement (Murthy, 1961).

The concepts of population parameter, estimator, and sample estimate are basic to estimation procedures. A parameter is a function of the population, and the mathematical expression of the function is its definition. The corresponding function of the sample is a statistic. An estimator is a rule or method, expressed in mathematical terms, for estimating the value of a parameter from sample observations. The sample estimate is the value which the estimator takes for a given sample. Parameters which will be discussed in the context of the population being sampled include the total, mean, and variance of the population (Murthy, 1961).

The next two sections will deal with the student and question populations. The discussion of each population will cover (i) the nature of the unit, population, and sample; (ii) parameters and estimators; and (iii) implications and relevance for psychological measurement and educational evaluation.

Sampling Populations of Students

A sampling procedure can be adopted when it is necessary to make inferences about all students belonging to a defined population but it is not possible to observe all of them. The population will be defined according to the purpose of the investigation. Thus, if the purpose is to investigate the mathematics proficiency of all school final students (e. g., Gayen *et al*, 1961) on a national or state scale, the population will be defined as all students completing the final year of schooling in the country or state. The unit is then a student completing the final year of school, and the sample will consist of final year students selected by a particular probability scheme.

A hypothetical example will be used in explaining the estimation procedures. In this example, the sample is to be selected by simple random sampling without replacement, and the characteristic of each unit, which is a student, is a mathematics achievement score denoted by X_h ($h = 1, 2,..., N$). Three parameters will be considered for this example: total, mean, and modified variance. The *total* is the aggregate of values obtained by summing the scores over all students. The *mean* is the expected score for any student in the population selected at random. The modified *variance* represents the variability of the scores for the population of students, and its square root is the standard deviation. The definitions and estimators for these parameters are given in Table 2.

Table 2

Estimation Notation for Simple Random Samling, without Replacement of a Population of Students

Sl No.	Population Parameter	Symbol	Definition	Sample Statistic	Symbol	Estimator
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	Total	X	$\sum_{h=1}^{N} X_h$	Total	x	$\frac{N}{n} \sum_{h=1}^{n} x_h$
2	Mean	$\bar{X}$	$\frac{1}{N} \sum_{h=1}^{N} X_h$	Mean	$\bar{x}$	$\frac{1}{n} \sum_{h=1}^{n} x_h$
3	Modified variance	S^2	$\frac{1}{(N-1)} \sum_{h=1}^{N} (x_h - X)^2$	Modified variance	s^2	$\frac{1}{(n-1)} \sum_{h=1}^{n} (x_h - \bar{x})^2$

A number of situations in psychological measurement and educational evaluation call for a statement of population performance. In psychological measurement, norms for aptitude and achievement tests are statements of the expected distribution of scores in the population. Yet it is hardly feasible to administer all such tests to all students in the population, usually defined as either a national or state population. Practical considerations suggest that a probability sample of the population could be tested and the expected distribution of scores estimated from the sample distribution. That this approach has not been seriously adopted has been pointed out by Lord (1962), who mentions that norms are frequently based on the performance of students from schools willing to cooperate with the particular test publisher. Comparative studies of any psychological measure, e.g.I.Q., in different populations assume that the observed values represent the whole population, yet the data are necessarily limited to samples only. Here again the application of statistical sampling theory would allow inferences about the population to be made from sample data. Studies of the item parameters, reliability, and validity of specific apttude and achievement tests would have greater generalizability if they were based on probability samples of the desired populations. In short, whenever an inference is to be made about the test performance of a population, but the entire population cannot be tested, statistical sampling theory can be usefully applied.

Educational evaluation by public examination is analogous to a census in that it requires observation of all units in the population. Thus, for every student appearing in a public examination, a mark sheet must be prepared. Detailed studies concerning educational standards, prediction of future success, and examination reform, to name a few topics, would usually be very expensive and time consuming if they were carried out on the entire population. Errors would be unassessable and unassessed. The collection of data and estimation therefrom by a statistical sampling procedure is, however, feasible in terms of cost and time, and in addition, the accuracy of the estimates can be determined. The work of Gayen *et al* (1961, 1963) illustrates the application of sampling theory to problems of educational evaluation.

In the investigation on achievement in English (Gayen *et al* 1963), the population was defined as all candidates for the School Final Examination in West Bengal in 1957, and the examination centres were treated as the strata into which the candidates were allocated. Then, using stratified sampling with proportional allocation among the different strata, a ten percent sample of candidates was drawn. For this sample, the mark sheets in the compulsory English paper were analyzed. The scope of the investigation may be indicated by the analyses which were carried out for each question: number of attempts, percentage of failures and passes, highest score and its frequency, mean score, standard deviation and coefficient of variation, discriminating power, difficulty values, reliability and validity. This type of information is of value for an appraisal of educational standards as well as the effectiveness of the different types of questions. While it would not be conceivable to carry out this type of analysis on the entire population of students, it would be feasible and worthwhile to carry it out on a sample of students from the population. Gayen's studies illustrate the application of statistical sampling theory to educational evaluation.

Sampling Universes of Questions

The setting of examination papers or construction of achievement or attainment tests involves implicitly the conceptualization of the universe or entire field of knowledge subsumed by the particular subject and the choice of specific topics to be covered within that universe or field of knowledge. In a sense, the specific topics chosen from a sample of the potentially available topics, and in the terminology of statistical sampling theory, the school or university syllabus can be viewed as a sampling frame. The process of choice is ordinarily governed by subjective preferences, biases and previous habits of the paper-setters, rather than by an objective scheme based on the laws of probability. The resulting question paper or attainment test lacks the advantages which would accrue to questions randomly selected from the universe of questions, among which may be mentioned freedom from examiner's bias and a measure of the accuracy in estimating the performance of individual students.

In sampling universes of questions, the term *unit* is used to refer to an examination question or a multiple-choice attainment question. As indicated in Table 1, the unit is designated by U_j. The *universe of questions* (Mahalanobis, 1960) is conceived of as the population of units, that is, the aggregate of examination questions or multiple-choice attainment questions. For each academic subject there can be a corresponding universe of questions, listed with proper identification in the sampling frame. The set of questions which forms the examination paper or attainment test is a *sample* of questions which has been selected from the universe of questions, as listed in the sampling frame according to the laws of probability. If the universe is regarded as homogeneous, simple random sampling of the questions is possible. But, as is more likely, if there are homogoneous topics within the universe, they can be used to stratify the universe of questions, and stratified sampling* is the more appropriate procedure.

The parameters and estimators of the universe of ques ions deal with the answers of any student to those questions (Das, 1965a, 1965b). For any student, E_h, the characteristic of the sampling unit, U_j is his answer The characteristic's value is denoted by Y ; the value may be quantitative (e. g , marks or points correct) or qualitative (correct or wrong). When stratified sampling is employed, the value is denoted by Y_{ij} in which i indicates stratum and j indicates unit within the stratum. The parameters of the universe of questions are regarded as the attributes of a student's answers to those questions. Restricting the present discussion to qualitative evaluation (Y_{ij} is 1 if correct or 0 if wrong), the *total* is the number of correct answers obtained by the student. The *mean* is a proportion which may be interpreted as the probability that any question in the universe would be correctly answered by the student. If the mean is multiplied by 100, it becomes the familiar percentage correct. The modified *variance* indicates the variability of the student's answers. In stratified sampling, the *sampling variance* of the mean is based on the variance within strata. The sampling variance indicates the precision of the mean estimator. Table 3 presents the parameters and estimators appropriate for a student's answers to a sample of questions selected by stratified sampling without replacement. In Table 3, the universe of questions is divided into K strata ($i = 1, 2, \ldots, K$) and the number of questions in the i^{th} stratum is denoted by M_i in the universe and m_i in the sample for $i = 1, 2, \quad, K$.

Table 3
Estimation Notation for Stratified Sampling, without Replacement, of the Universe of Questions

Sl. No.	Universe: Parameter	Universe: Symbol	Universe: Definition	Sample: Statistic	Sample: Symbol	Sample: Estimator
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	Total	Y	$\sum_{i=1}^{K}\sum_{j=1}^{M_i} Y_{ij}$	total	y	$\sum_{i=1}^{K} \frac{M_i}{m_i} \sum_{j=1}^{m_i} y_{ij}$
2	Mean	$\bar{Y}$	$\frac{1}{\bar{M}} \sum_{i=1}^{K}\sum_{j=1}^{M_i} Y_{ij} = \sum_{i=1}^{K} \frac{M_i}{M} \bar{Y}_i$	mean	$\bar{y}$	$\sum_{i=1}^{K} \frac{M_i}{M} \bar{y}_i$
			where $\bar{Y}_i = \frac{1}{M_i} \sum_{j=1}^{M_i} Y_{ij}$			where $\bar{y} = \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij}$
3	Modified Variance Within Strata	S_i^2	$\frac{1}{(M_i - 1)} \sum_{j=1}^{M_i} (Y_{ij} - \bar{Y}_i)^2$	modified variance within strata	s_i^2	$\frac{1}{(m_i - 1)} \sum_{j=1}^{m_i} (Y_{ij} - \bar{y}_i)^2$
4	Sampling Variance of Sample Mean	$V(\bar{y})$	$\sum_{i=1}^{K} (\frac{M_i}{M})^2 (\frac{1}{m_i} - \frac{1}{M_i}) S_i^2$	sampling variance of sample mean	$\hat{V}(\bar{y})$	$\sum_{i=1}^{K} (\frac{M_i}{M})^2 (\frac{1}{m_i} - \frac{1}{M_i}) s_i^2$

Having estimated the population mean from the sample mean and its sampling variance from a single sample of questions, the possibilities arising from multiple samples of questions will be considered. If the population mean $\bar{Y}$ is expressed as a percentage

$$P = 100\ Y \qquad [1]$$

and the sample mean $\bar{y}$ is similarly expressed

$$p = 100\ \bar{y}, \qquad [2]$$

p serves to estimate the true percentage, P, of correct answers to the universe of questions. If the number of multiple samples is denoted by r, and the u_{th} sample estimate of P is given by p_u (u=1, 2,......, r), the combined estimate of P is

$$\bar{p} = \frac{1}{r} \sum_{u=1}^{r} p_u \qquad [3]$$

and the estimator of its sampling variance is

$$\hat{V}(\bar{p}) = \frac{1}{r(r-1)} \sum_{u=1}^{r} (p_u - \bar{p})^2 \qquad [4]$$

(Murthy, 1961, Page 227).

It is possible to obtain the confidence interval for P assuming that $\bar{p}$ is normally distributed with mean P and variance $V(\bar{p})$. It follows that the statistic

$$\frac{\bar{p} - P}{\sqrt{\hat{v}(\bar{p})}}$$

is distributed as Student's t with (r-1) degrees of freedom. Then, the 100 (1-α)% confidence limits are given by

$$\bar{p} - t_{\alpha, r-1} \sqrt{\hat{v}(\bar{p})},\ \bar{p} + t_{\alpha, r-1} \sqrt{\hat{v}(\bar{p})}, \qquad [5]$$

the length of the confidence interval is given by

$$2(t_{\alpha, r-1} \sqrt{\hat{v}(\bar{p})} \qquad [6]$$

and the confidence interval is then

$$\bar{p} \pm t_{\alpha, r-1} \sqrt{\hat{v}\ \bar{p})} \qquad [7]$$

For 95% confidedence, α is taken at. 05, and for 99% confidence α is taken at .01.The value of $t_{\alpha, r-1}$ is read frcm the tables at the desired level of (1-α/2) and required degrees of freedm (Rao, Mitra, and Matthai,1966). The student's true knowledge, i. e., true percentage correct answers to the universe of questions, is expected to lie, with 100 (1—α)% confidence (usually 95% or 99%), within these limits. The longer the interval length (see [6]) the less accurate the estimate, and the shorter the interval length, the more accurate the estimate. The length of the confidence interval provides a measure of the accuracy of the estimate which is individually determined. This approach may be contrastcd with the standard error of measurement commonly employed in mental test theory, which imposes an estimate of the group's error on the individual student (e. g., Gulliksen, 1950). To illustrate this approach, Table 4 gives confidence intervals along with $\bar{p}$, the estimate of P, for five students, three degrees of freedom for each student.

Table 4

Confidence Limits and Probabilities for Individual Students

Sl. No.	Function: Name	Function: Symbol	Equation Number	Student 1	Student 2	Student 3	Student 4	Student 5
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1	Estimate of P, sample 1	p_1	[2]	53.33	23.33	60.00	56.67	30.00
2	Estimate of P, sample 2	p_2	[2]	56.67	23.33	36.67	56.67	43.33
3	Estimate of P, sample 3	p_3	[2]	63.33	33.33	56.67	66.67	46.67
4	Estimate of P, sample 4	p_4	[2]	50.00	43.33	63.33	63.33	46.67
5	Mean of sample estimates	$\bar{p}$	[3]	55.83	30.83	54.17	60.84	41.67
6	Square root of estimate of sampling variance of $\bar{p}$	$\sqrt{\hat{V}(\bar{p})}$	[4]	2.85	4.79	5.99	2.50	3.97
7	Length of 95% confidence interval of P	$2[3.18\sqrt{\hat{V}(\bar{p})}]$	[6]	18.12	30.46	38.10	15.90	25.24
8	95% confidence limits of P	$\bar{p} \pm 3.18\sqrt{\hat{V}(\bar{p})}$	[7]	64.89 46.77	46.06 15.60	73.22 35.12	68.79 52.89	54.29 29.05
9	Length of 99% confidence interval of P	$2[5.84\sqrt{\hat{V}(\bar{p})}]$	[6]	33.28	55.94	69.96	29.20	46.36
10	99% confidence limits of P	$\bar{p} \pm 5.84\sqrt{\hat{V}(\bar{p})}$	[7]	72.47 39.19	58.80 2.86	89.15 19.19	75.44 49.24	64.85 18.49
11	Probability, first class	$\hat{\varepsilon}_1$	[13]	.1422	.0046	.2172	.6743	.0103
12	Probability, second class	$\hat{\varepsilon}_2$	[14]	.8265	.0151	.6069	.3191	.1050
13	Probability, third class	$\hat{\varepsilon}_3$	[15]	.0312	.5884	.1627	.0065	.8581
14	Probability, failure	$\hat{\varepsilon}_4$	[16]	.0001	.3919	.0132	.0001	.0266
15	Sum of probabilities	$\hat{\varepsilon}_1+\hat{\varepsilon}_2+\hat{\varepsilon}_3+\hat{\varepsilon}_4$	[17]	1.000	1.000	1.000	1.000	1.000

It is also possible to estimate the probability with which a student is placed in the "first class", "second class" or "third class", or classified as 'fail'. In terms of three constants, called d^1, $d2$, and $d3$ four individual probabilities are defined :

ε_1 = prob. ($\bar{p} \geq d_1$), first class; [8]

ε_2 = prob. ($d_1 > \bar{p} \geq d_2$), second class; [9]

ε_3 = prob. ($d^2 > \bar{p} \geq d_3$), third class; and [1]

ε_4 = prob. ($0 \leq \bar{p} < d_3$), failure [11]

Assuming that $\bar{p}$ is normally distributed with mean P and variance $V(\bar{p})$, these individual probabilities can be estimated by

$$\hat{\varepsilon}_1 = T_{r-1}\left(\frac{d_1 - \bar{P}}{\sqrt{\hat{v}(\bar{p})}}\right), \quad [12]$$

$$\hat{\varepsilon}_2 = T_{r-1}\left(\frac{d_2 - \bar{P}}{\sqrt{\hat{v}(\bar{p})}}\right) - T_{r-1}\left(\frac{d_1 - \bar{P}}{\sqrt{\hat{v}(\bar{p})}}\right), \quad [13]$$

$$\hat{\varepsilon}_3 = T_{r-1}\left(\frac{d_3 - \bar{P}}{\sqrt{\hat{v}(\bar{p})}}\right) - T_{r-1}\left(\frac{d_2 - \bar{P}}{\sqrt{\hat{v}(\bar{p})}}\right), \quad [14]$$

and

$$\hat{\varepsilon}_4 = 1 - T_{r-1}\left(\frac{d_3 - \bar{P}}{\sqrt{\hat{v}(\bar{p})}}\right), \quad [15]$$

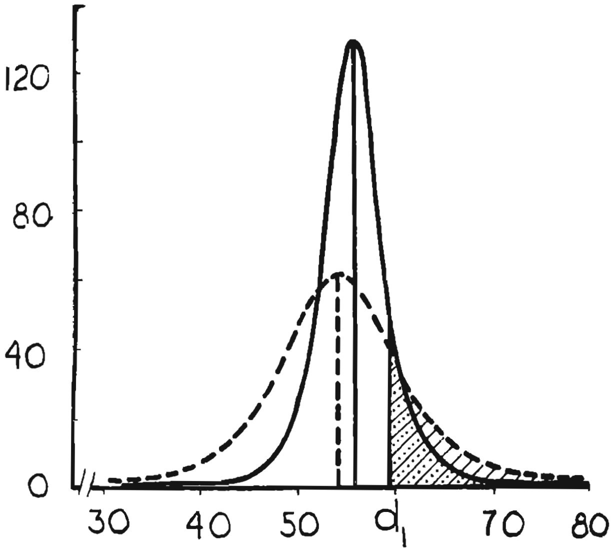
It can be seen that

$$\hat{\varepsilon}_1 + \hat{\varepsilon}_2 + \hat{\varepsilon}_3 + \hat{\varepsilon}_4 = 1. \quad [16]$$

According to [12] to [15] the difference between any constant $\bar{d}$ and $\bar{p}$, divided by $\sqrt{\hat{v}(\bar{p})}$, is interpreted as t with (r—1) degrees of freedom and accordingly T_{r-1} is the associated area under the t curve to the right of that value. It may be obtained by reference to tables of the probability integral of the t-distribution, subtracting the tabled value from unity (Pearson and Hartley, 1956). For a standard university examination curve with u=45 and o=15, (University Grants Commission 1962, page 103), the constants have the following values : d_1=59.5, d_2=?7.5, and d^3= 29.5, Using these values, the probabilities of being placed in each of the possible classes have been computed for the students whose confidence intervals are given in Table 4. The probabilities are shown in rows sl. no. 11 to 14 of the same table. The confidence intervals and probabilities for two students with nearly the same value of $\bar{p}$ but different values of $\sqrt{\hat{v}(\bar{p})}$ are shown by Figure 1. (Note: for negative values of t, use the tabled value corresponding to a positive t but do not subtract the tabled value from unity.

Both from a theoretical and practical point of view, construction of tests and setting of examination papers by sampling universes of questions would have interesting implications. In considering the theoretical implications, it should be recalled that classical psychometric theory is not based on theory of probability, but rather upon the concepts of correlation and true score. As a result, there is no basis for generalizing from the the actual questions studied to the potential set of all questions. The objective, however, of psychometric measurements is to make generalizations about the attributes of the potential set of all questions. Sampling theory can provide a statistical foundation for psychological measurment which will allow inferences to be made about the potential set of all questions, that is, the universe of questions. While Lord (1955, 1959), Cronbach and Azuma (1962), Cronbach, Rajaratnam, and Gleser (1963), Cronbach, Schonemann, and Mckie

Figure 1



Legend

Figure 1, 'Student's t' curves for two students: A, $\bar{P}$ = 55.83, $\sqrt{V(\bar{p})}$=2.85, solid line; B, $\bar{P}$= 54.17, $\sqrt{\hat{V}(\bar{p})}$ =5.99, dashes.

The abscissa represents percentage right score. The ordinate gives the value of the frequency function, divided by $\sqrt{\hat{V}(\bar{p})}$ for 'Student's t' at 3 degrees of freedom. Vertical lines are erected at $\bar{P}$ for A (solid line) and B (dashes), and at dl = 59.5. The area under each curve is unity and the area to the right of d_1 represents the probability of being placed in the first class. The areas to the right of d_1 are stippled for A and hatched for B. The two students were studying in the M. Sc. Class in Zoology, and their knowledge was assessed by four samples from the universe of "Questions on Biology" (Appraisal Division, Indian Statistical Institute).

(1965) and Rajaratnam, Cronbach, and Gleser (1965), have considered reliability of tests in the context of sampling theory, they have not dealt with estimation of individual performance, nor have they constructed and sampled a question universe. Lord (1964) has discussed the concept of true score in terms of sampling theory, while the discussion in the preceding two paragraphs has been concerned with the estimate of the true score of an individual student and its confidence interval, utilizing sampling theory (Das,1965a;1965b). The work described in this paper is possibly the first attempt to actually sample an actual universe of questions, to administer the samples, and carry out the described estimation procedures.

From a practical point of view, sampling theory provides for rapid selection of questions according to requirements (Mahalanobis, 1960). Given the desired universe of questions, stratified in terms of topic and difficulty, it is possible to select samples of questions according to specification of difficulty and subject coverage. If equivalent sets of questions are required, so as to obviate coping among candidates in an examination hall, they can be readily obtained by sampling the universe of questions. Sampling theory provides a method whereby standards over space (geographical and linguistic regions) and time (year to year) can be maintained. Being objective and quantititative in nature, sampling of questions can be carried out by electronic computers, and hence enjoys the speed and accuracy of modern data processing equipment. The sampling approach can be applied at all examination levels, e. g., school final, degree, postgraduate and in large-scale high-level selection, e. g., for the administrative services. The serious development of universes of questions in different languages and indifferent subjects has been taken up, for the first time in India or abroad, in the Indian Statistical Institute(Mahalnobis, 1960; Das, 1965a, 1965b)

Discussion

Briefly stated, sampling deals with selection of units from an aggregate, and estimation is concerned with the values of the aggregate. For the population of students, the unit is a student, and the value has been taken as that student's total score or marks For the population of questions, the unit is a question, and the value has been taken as a single student's scored answer. It is possible to place the sampling and estimation for these two populations in a combined statistical framework. If the two populations are considered simultaneonsly, they may be represented by a matrix, in which the columns represent questions and the rows represent students (Lord, 1955, 1959; Wilks, 1962). Table 5 presents the student-question matrix for the population (columns (2) tu (5)) and for the sample (columns (9) to (12)) for the simplest case in which the two populations are unstratified. These matrices can be viewed in at least five different ways, peseribed by the the following situations.

Table 5

Population and Sample of Matrices Representing the Values of Students' Answers to Questions

Population							Sample						
Students	Questions				Row		Students	Questions				Row	
	1	2	j	M	Total	Mean		1	2	j	m	Total	Means
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
1	Y_{11}	Y_{12}	Y_{1j}	Y_{1M}	$Y_{1.} = X_1$	$\bar{Y}_{1.}$	1	y_{11}	y_{12}	y_{1j}	y_{1m}	$y_{1.} = x_1$	$\bar{y}_{1.}$
2	Y_{21}	Y_{22}	Y_{2j}	Y_{2M}	$Y_{2.} = X_2$	$\bar{Y}_2$	2	y_{21}	y_{22}	y_{2j}	y_{2m}	$y_{2.} = x_2$	$\bar{y}_{2.}$
.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.
h	Y_{h1}	Y_{h2}	Y_{hj}	Y_{hM}	$Y_{h.} = X_h$	$\bar{Y}_{h.}$	h	y_{h1}	y_{h2}	y_{hj}	y_{hm}	$y_{h.} = x_h$	$\bar{y}_{h.}$
.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.
N	Y_{N1}	Y_{N2}	Y_{Nj}	Y_{NM}	$Y_{N.} = X_N$	$\bar{Y}_{N.}$	n	y_{n1}	y_{n2}	y_{nj}	y_{nm}	$y_{n.} = x_n$	$\bar{y}_{.}$
Column Total	$Y_{.1}$	$Y_{.2}$	$Y_{.j}$	$Y_{.M}$	$Y_{..} = X$	$\sum_{h=1}^{N} \bar{Y}_{h.}$	column total	$y_{.1}$	$y_{.2}$	$y_{.j}$	$y_{.m}$	$y_{..} = x$	$\sum_{h=1}^{n} \bar{y}_{h.}$
Column mean	$\bar{Y}_{.1}$	$\bar{Y}_{.2}$	$\bar{Y}_{.j}$	$\bar{Y}_{.M}$	$\frac{Y_{..}}{N} = \bar{X}$	$\bar{Y}_{..}$	column mean	$\bar{y}_{.1}$	$\bar{y}_{.2}$	$\bar{y}_{.j}$	$\bar{y}_{.m}$	$\frac{y_{.}}{n} = \bar{x}$	$\bar{y}_{..}$

Situation (i) : The notation for dealing with sampling of students has been given in Table 1; the units of the student population are also denoted in column (1) of Table 5. Estimation notation for the student population has been presented in in Table 2; the score X_h is also shown in column (6) of Table 5. In other words, the population of students is symbolically represented in column (1) of Table 5, and the score X_h for the h^{th} student is obtained by summing over columns (2) to (5), i. e. , questions 1 to M, The latter procedure implies that each student has answered the universe of questions; it is more likely, in fact, that the score is based on a sample of questions, hence x_h as denoted in column (14) of Table 5. It should be emphasized that X_h as used in Table 2 is not based on a random sample of questions while Table 5 shows that it can be so obtained. This difference has the further implication that generalization from x_h to X_h is not possible unless the m questions are a random sample from the universe of questions.

Situation (ii). The notation for sampling the universe of questions is summarized in Table 1. The units of the universe and sample of questions are also shown by columns (2) to (5) and (9) to (12) respectively of Table 5. As dealt with in Table 3, estimation for the universe of questions has been restricted to the evaluation of the perfomance of individal students, taken singly. Thus, the population values for the student's knowledge are represented by row h, columns (2) to (5), in Table 5. The sample values for the same student are presented in row h, columns (9) to (12). Table V also gives the population total $Y_{h.}$ in column (6), the population mean $\bar{Y}_{h.}$ in column (7), the sample total $y_{h.}$ in column 13, and the sample mean $\bar{y}_{h.}$ in column 14. The parameters and statistce of Table 3 apply to row h of Table 5.

Situation (iii). Table 5 also permits a psychometric analysis of the questions. The column means of Table 5 give the difficulty values, as proportion of the population (columns (2) to (5)) or sample (columns (9) to (12)) correctly answering the questions. The true difficulty values are those of the population, while the estimates are those of the sample. It may be noted that the columns of Table 5 are considered separately in this psychometric analysis, in contradistinction to situation (ii) in which the rows are considered separately. Statistical generalization is from the sample to the population of students. While random sampling from the population of students is obviously desirable in psychometric analyses, it is unfortunately rather infrequent.

Situation (iv). Column (4) provides the array of true scores to question u_i and from it, not only the difficulty (the column mean) but also discrimination and other psychometric item parameters can be obtained (Das, 1964). The corresponding sample statistics can be obtained from column (11) for question v_j. As in situation (iii), the questions are considered separately.

Situation (v). Thus far, four situations or types of data have been derived from the matrices of Table 5;

(i) row mean, as a scalar which is a student index;

(ii) any row as a vector, which is an array of a student's scored answers;

(iii) column mean, as a scalar which is a question index; and

(iv) any column, a vector, which is an array of scored answers to a question.

Simultaneous consideration of the two populations permits definition of a random variable which represents question performance as a function of the population tested. Denoted Y_{hj} (h=1, 2, ..., N; j=1, 2,......, M) n the population (column (4), row h of Table (5) and y_{hj} in the sample (column (11), row h of Table 5) in the present notation, it is designated x (u, v) by Wilks (1962). For any question chosen at random from the designated universe of questions, and for any student chosen at random from the specified population of students, Y_h is a random variable which takes the value of 0 or 1. The sample statisti may be interpreted as the estimated probability that the question will be answered correctly, i. e., $Y_{hj}=1$. Obtaining the value Y_h^j would be particularly appropriate for comparing matrics. Holding the universe of questions constant, the value of y_{hj} for samples from different populations of students could be compared. Conversely, keeping the student population constant, y_h for samples of questions from different universes could be compared. Finally, a set of question universes and a set of student populations could be chosen which would represent successive levels of education and achievement standard; yhj could serve as an index of the changes in educational attainment which occur with maturation and devolopment.

The matrix of students and questions has been discussed in terms of the psychometric applications described in situations (iii) and (iv) above by Lord (1955, 1959, 1964). The matrix has also been conceptualized, though not expressed in formal statistical terminology, by Brunswik (1956), Guttamn (1950), and Thomson (1956). Conceiving of a universe of social objects or social stimuli, Brunswik (1956) clearly indicated the difference in generalization from sampling—of—objects and sampling—of—subjects. His sampling—of—objects corresponds to the sampling of questions, and his sampling—of—subjects to the sampling of students. Brunswik applied these concepts to social psychology and the psychology of preception. Guttamn (1950) proposed the concept of an attitude universe or universe of attributes, for which the attitude scale or questionnaire served as a sample. He introduced the ideas of scalability and reproducibility to evaluate the adequacy of the sample, and also clearly distinguished between the sampling of attiudes and the sampling of individuals. Thomson (1956) suggested that tests sample the mental bonds existing in the brain, and recognized the differences arising from sampling of questions and sampling of individuals. The theoretical formulations of Brunswik, Guttamn, and Thomson share certain features; among these are the conceptualization of a universe of objects to be sampled, the distinction between generalization from subject populations and from object populations, and the absence of sampling and estimation theory. The sampling theory appropriate for the universe of questions which has been proposed in this paper could readily be extended to cover the formulations of Brunswik, Guttamn, and in so doing criticisms of their work raised by Campbell and Kerckhoff (1957) and Loevinger (1965), would be met.

REFERENCE

Brunswik, E. (1956) *Perception and the representative design of psychological experiments.* Berkely and Los Angles, University of California Press.

Campbell, E. Q. and Kerckhoff, A. C. (1957) A critique of the concept "Universe of attributers". *Publ. Opin. Quart.* 3: 295—803.

Cronbach, L. J. and Azuma, H. (1962) Internal consistency reliability formulas applied to randomly sampled single-factor tests: an empirical comparison. *Educ. Psychol. Measur.* 23 : 645–665.

Cronbach. L. J., Rajaratnam, N., and Gleser, G. C. (1963) Theory of generalizability: A liberalization of reliability theory. *J. Stat. Psychol.* 16 : 137—163.

Cronbach, L; J., Schonemann, P. and McKie, D. (1965) Alpha coefficients for stratified--parallel tests. *Educ. Psychol. Measur.* 25 : 291–312.

Das, R. S. (1964) Items analysis by probit and fractile graphical methods. *Brit. J. Stat. Psychol.* 17: 51—64.

Das, R.S. (1965) Prediction of individual performance in production and management. (To be published in the Proceedings of the Seminar on Behavioral Sciences in Management, Indian Institute of Management, Calcutta, August 1965) (a)

Das, R. S. (1965) Some models for assessment of intelligence and scholastic attainment. *J. Psychol. Res.* (In press).

Gayen, A. K., Nanda. P. B., Mathur, R. K., Dubey, S.D., Duari, P. and Bhattacharya, N. (1961) *Measurement of achievement in mathematics.* Ministry of Education, Govt. of India, New Delhi.

Gayen, A. K., Nanda, P. B., Bhattacharya, N., Mukerjee, H. N., Mathur, R. K. (1963) *Measurement of achievement in English.* National Council of Educational Research and Training, New Delhi.

Gulliksen, H. (1950) *Theory of mental tests.* New York. Wiley.

Guttaman, L. (1950) The basis for scalogram analysis. Chapter 3 in S. A. Stouffer (Ed), *Measurement and prediction.* Princeton, University Press.

Loevinger, J. (1965) Person and population as psychometric concepts. *Psychol. Rev.* 72 : 143-155.

Lord, F. M. (1955) Sampling fluctuations resulting from the sampling of test items. *Psychometrika.* 20 : 1–22.

Lord, F. M. (1959) An approach to mental test theory. *Psychometrika.* 24 : 283—302.

Lord, F. M. (1962) Estimating norms by item sampling. *Educ. Psychol. Measur.* 22: 259-267.

Lord, F. M. (1964) Nominally and rigorously parallel test froms. *Psychometrika.* 29: 335-345

Mahalanobis, P. C. (1960) Next steps in planning. *Sankhya.* 22: 143—172.

Murthy, M. N. (1961) *Introduction to sampling theory.* (cyclostyled) Calcutta, Indian Statistical Institute.

Pearson, E.S. and Hartley, H. O. (Eds.) (1956) *Biometrika tables for statisticians.* Volume I, Fourth Edition. Cambridge, University Press.

Rajaratnam, N., Cronbach, L. J., and Gleser, G. C. (1965) Generalizability of stratified-parallel tests. *Psychometrika.* 30 : 39–56.

Rao, C. R., Mitra, S. K., and Matthai, A. (1966). *Formulae and tables for statistical work.* Calcutta, Statistical Publishing Society.

Thomson, G. (1956). *The factorial analysis of human ability.* Fifth edition, first reprint. London, Univesity of London Press.

Wilks, S. S. (1962). *Mathematical statistics.* New York, John Wiley.

University GrantsCommission (1962). *Report on examination reform.* New Delhi.

Rhea S. Das

Appraisal Division
Indian Statististcal Institute
20?, Barrackpare Trnuk Road
CALCUTTA-35.

★