CHAPTER 17

SOME LATENT TRAIT MODELS

17.1 Introduction

In this chapter we shall consider in detail several models of tests, some of which have been introduced more briefly above (Sections 15.6 and 16.1 through 16.5). We shall now describe these models in self-contained mathematical terms to prepare ourselves to examine them, subsequently, in relation to theories and applications of tests. These models have been developed primarily in connection with tests of various general or special abilities, although it has proved of interest to consider them also in relation to the study of other kinds of traits, such as attitudes. For convenience, we shall refer to the trait in question simply as "ability".

We consider here tests consisting of items each to be scored 0 or 1, with u_g as the generic symbol for the score on item g and with $\mathbf{v}' = (u_1, \ldots, u_g, \ldots, u_n)$ representing the set of scores, or the *response-pattern*, on a test of n items. This notation tacitly refers to scores of some one individual subject; when necessary, scores of a subject indexed a can be denoted more explicitly by $\mathbf{v}'_a = (u_{1a}, \ldots, u_{ga}, \ldots, u_{na})$.

Item scores u_g are related to an ability θ by functions that give the probability of each possible score on an item for a randomly selected examinee of given ability. These functions are

$$Q_g(\theta) = \text{Prob}\,(U_g = 0 \mid \theta)$$

and the *item characteristic curve* (*ICC*)

$$P_g(\theta) = \text{Prob}\,(U_g = 1 \mid \theta) = 1 - Q_g(\theta).$$

These formulas are conveniently combined in the probability distribution function of U_g:

$$f_g(u_g \mid \theta) \equiv \text{Prob}\,(U_g = u_g \mid \theta) = P_g(\theta)^{u_g} Q_g(\theta)^{1-u_g} \equiv \begin{cases} P_g(\theta) & \text{if} \quad u_g = 1, \\ Q_g(\theta) & \text{if} \quad u_g = 0, \end{cases}$$

where f_g is defined in a persons or in a persons-by-replications space.

We note that the regression function of any item response u_g is identical with its item characteristic curve since

$$\mathscr{E}(U_g \mid \theta) = 1 \cdot f_g(1 \mid \theta) + 0 \cdot f_g(0 \mid \theta) = P_g(\theta).$$

Any item for which $P_g(\theta)$ has a constant value independent of θ is not an indicant (and hence *a fortiori* not a measure) of θ in the sense of Section 1.4. In most cases of interest here, we shall have $P_g(\theta)$ strictly increasing in θ, so that u_g will be an indicant and a measure of θ. We do not assume a probability distribution for θ in any part of the present treatment of this subject. For an extension of the theory which makes use of this assumption, the reader should see Birnbaum (1967).

These functions do not determine unequivocally the relation between an ability and a complete response pattern $\mathbf{v}' = (u_1, \ldots, u_n)$ unless they are supplemented in some definite way. The additional assumption found most useful in test theory and its applications, as well as the simplest assumption mathematically, is *local independence* (see Section 16.3). This assumption implies the mathematical condition of *statistical independence between responses* by a subject to different items; it is represented by the usual probability product form

$$\begin{aligned}\text{Prob}\,(\mathbf{V} = \mathbf{v} \mid \theta) &\equiv \text{Prob}\,(U_1 = u_1, \ldots, U_n = u_n \mid \theta)\\ &= \text{Prob}\,(U_1 = u_1 \mid \theta) \cdots \text{Prob}\,(U_n = u_n \mid \theta)\\ &= \prod_{g=1}^{n} P_g(\theta)^{u_g} Q_g(\theta)^{1-u_g}.\end{aligned}$$

For example, the product form $\text{Prob}\,[(U_1, U_2) = (1, 1) \mid \theta] = P_1(\theta)P_2(\theta)$ represents the fact that any subject of ability θ gives independent responses to items 1 and 2; that is, that the probability $P_2(\theta)$ of his correctly answering item 2 is the same as the conditional probability of his correctly answering item 2, given that he has correctly answered item 1. The relations of this assumption to more general models and theories, in which several abilities are considered jointly, have already been discussed in Section 16.2.

One basic aspect of the questions of validity and empirical and theoretical content discussed in Chapter 1, as they apply to the models introduced here, may be illustrated conveniently at this point. Consider any item, and consider a series of groups of subjects in which each subject is assumed to have common ability. Suppose that the probabilities of correct responses to the item in the respective groups are p_i, where $0 \leq p_1 < p_2 < \cdots < p_m \leq 1$. Since we have mentioned all the empirically meaningful aspects of a model of a single item, we still remain free to choose arbitrarily a series of numbers θ_i,

$$-\infty < \theta_1 < \theta_2 < \cdots < \theta_m < \infty,$$

which we may call the *true ability scores* of the respective groups. The choice of these numbers θ_i amounts to a choice of the specific form of an ICC function that shall represent the first item, since we *define* the function $P_1(\theta)$ as the correspondence between respective ability scores θ_i and values $P_i = P_1(\theta_i)$. Equivalently, given the numbers p_i, we can adopt *any* increasing function $P_1(\theta)$ as the ICC of the item: This choice associates an ability score θ_i, determined by $p_i = P_1(\theta_i)$, with the group of subjects scoring p_i.

These comments illustrate the fact that an essentially conventional element exists in the relations between ability levels θ and observable item responses. Once any specific strictly increasing form has been adopted for $P_1(\theta)$, for example, $P_1(\theta) = \Phi(2\theta - 1)$, the statement that a subject has ability $\theta = 2.1$ has empirical content and consequences in the contexts of models discussed here. For any second item (assuming local independence), the item characteristic curve $P_2(\theta)$ has a value at $\theta = 2.1$ which is estimable from empirical data in the same sense as is $P_1(2.1)$. Thus we are *not free* to adopt by definition *any* number as the value of $P_2(2.1)$. Similarly we are not free to adopt *any* assumption restricting even partially the possible functional forms of any other item characteristic curves $P_g(\theta)$, $g = 2, 3, \ldots, n$. This illustrates the fact that in general it is empirically meaningful (nontautological) to assume that any specific model, or even any class of models of partially restricted form, is valid in relation to a specified population of items. Therefore it is possibly false and hence is subject to empirical confirmation (or partial confirmation or disconfirmation). On the other hand, the assumption that any chosen *single* item has an item characteristic curve of a specified functional form $P_g(\theta)$ that depends on ability θ is, when considered *in isolation*, acceptable in principle as a definition of the ability scale of θ values and is not an empirical specification.

17.2 The Logistic Test Model

A function which very nearly coincides with the normal ogive model treated in Section 16.5, and which has advantages of mathematical convenience in several areas of application, is the logistic (cumulative) distribution function

$$\Psi(x) = e^x/(1 + e^x) \equiv 1/(1 + e^{-x}), \qquad -\infty < x < \infty. \qquad (17.2.1)$$

The inverse function is $x = \log[\Psi/(1 - \Psi)]$. For simple descriptive purposes, any graph of a cumulative normal distribution function $\Phi(x)$ would serve equally well to illustrate this function, since it has been shown (Haley, 1952, p. 7) that

$$|\Phi(x) - \Psi[(1.7)x]| < 0.01 \qquad \text{for all } x. \qquad (17.2.2)$$

We may state this relation in another way: The logistic cdf $\Psi(x)$ differs by less than 0.01, uniformly in x, from the normal cdf with mean zero and standard deviation 1.7; that is,

$$|\Phi(x/1.7) - \Psi(x)| < 0.01 \qquad \text{for all } x.$$

The probability density function (pdf) corresponding to the logistic cdf is

$$\psi(x) = e^{-x}/(1 + e^{-x})^2 \equiv \Psi(x)[1 - \Psi(x)] \equiv \tanh^{-1}(x). \tag{17.2.3}$$

Berkson (1957) has given detailed tables of $\Psi(x)$ and $\psi(x)$. Of course, tables of the exponential function and of the hyperbolic tangent are also available, and hence direct computation of values of these functions is not difficult.

The *logistic test model* is determined by assuming that item characteristic curves have the form of a logistic cumulative distribution function:

$$P_g(\theta) = \Psi[DL_g(\theta)] \equiv [1 + e^{-DL_g(\theta)}]^{-1} = [1 + e^{-Da_g(\theta - b_g)}]^{-1}, \tag{17.2.4}$$

where $L_g(\theta) = a_g(\theta - b_g)$, and $g = 1, 2, \ldots, n$. We have also

$$Q_g(\theta) = 1 - \Psi[DL_g(\theta)] \equiv [1 + e^{DL_g(\theta)}]^{-1},$$

$$P_g(\theta)/Q_g(\theta) = e^{DL_g(\theta)}, \qquad \text{and} \qquad \frac{\partial}{\partial\theta}P_g(\theta) = Da_gP_g(\theta)Q_g(\theta).$$

(Again, we do not interpret $P_g(\theta)$ here as a probability distribution function, even when it has the mathematical properties of one.) Here a_g and b_g are item parameters whose roles are generally the same as those of the item parameters in the normal ogive model because of the qualitative, and nearly exact quantitative, similarity between the models. The symbol D denotes a number that serves, at our convenience, as a unit scaling factor. To maximize agreement between quantitative details in the normal and logistic models, we can and usually shall take $D = 1.7$; then

$$P_g(\theta) = \Psi[1.7a_g(\theta - b_g)] \equiv (1 + e^{-1.7a_g(\theta - b_g)})^{-1}. \tag{17.2.4a}$$

For notational convenience, however, we shall often write the logistic model using the symbol D for the number 1.7.

We may view the logistic form for an item characteristic curve as a mathematically convenient, close approximation to the classical normal form, introduced to help solve or to avoid some mathematical or theoretical problems that arise with the normal model. Or we may view it as the form of a test model that is of equal intrinsic interest and of very similar mathematical form. The important questions of the validity of such models in observational and theoretical contexts are discussed elsewhere (see Sections 16.1 and 17.10).

The probability distribution function of a response u_g in a logistic test model is

$$\begin{aligned} f_g(u_g \mid \theta) &\equiv P_g(\theta)^{u_g}Q_g(\theta)^{1-u_g} \\ &\equiv Q_g(\theta)[P_g(\theta)/Q_g(\theta)]^{u_g} && (17.2.5) \\ &= \frac{\exp[Da_g(\theta - b_g)u_g]}{1 + \exp[Da_g(\theta - b_g)]}, && (17.2.6) \end{aligned}$$

and, under the assumption of local independence, the probability distribution function of a response pattern $\mathbf{v}' = (u_1, \ldots, u_n)$ is

$$\begin{aligned}\operatorname{Prob}(\mathbf{V} = \mathbf{v} \mid \theta) &= \prod_{g=1}^{n} f_g(u_g \mid \theta) = \prod_{g=1}^{n} Q_g(\theta) \prod_{h=1}^{n} \exp\,[Da_h(\theta - b_h)u_h] \\ &= \left[\prod_{g=1}^{n} Q_g(\theta)\right]\left[\exp\left(\theta D \sum_{g=1}^{n} a_g u_g\right)\right]\left[\exp\left(-D \sum_{g=1}^{n} a_g b_g u_g\right)\right].\end{aligned} \tag{17.2.7}$$

The principal features of mathematical simplicity that characterize the logistic test model are, as we shall see, implicit in this last form. In particular, "all the information about θ available in a response pattern $\mathbf{v}$" (in a sense to be specified) is given by the particular test score formula

$$x = x(\mathbf{v}) = \sum_{g=1}^{n} a_g u_g,$$

which does not depend on the difficulty parameters b_g. We may further illustrate the roles of item parameters and the properties of such a test score formula by considering an artificial example of the logistic test model. Let us take just four of the items whose parameters have the values represented in Fig. 16.5.1, namely, $g = 3, 4, 5$, and 6. (The same figure serves equally well here to illustrate either logistic or normal item characteristic curves.) We have $a_3 = 100$, $a_4 = 100$, $a_5 = 1$, and $a_6 = 1$. The test score is then

$$\begin{aligned} x &= 100y_3 + 100y_4 + y_5 + y_6 \\ &= 100(y_3 + y_4) + (y_5 + y_6). \end{aligned}$$

The possible values of x are just

$$\begin{array}{ccc} 0 & 1 & 2 \\ 100 & 101 & 102 \\ 200 & 201 & 202. \end{array}$$

We see that the major part of this ordering of subjects' response patterns, which is represented by the rows of the preceding array, is determined by the heavily weighted responses to the informative items y_3 and y_4. The only role of the less informative items in this example is to give a finer ordering compatible with the initial rough ordering. This example is extreme: Typical tests one meets in practice have more items and less extreme variation in weights a_g. With more nearly typical tests, it is usually possible to reverse an ordering of two response patterns based only on responses to several items if all items are taken into account in a suitable weighted composite score.

Table 17.3.1

Standard deviation of sample item-test biserials*

Test	Number of items	Sample item-test biserials	
		Mean	Standard deviation
Listening Comprehension	50	0.51	0.12
English Structure	70	0.48	0.11
Vocabulary	60	0.55	0.09
Reading Comprehension	30	0.54	0.09
Writing Ability	60	0.44	0.11

* From an internal Educational Testing Service report (SR–66–80) prepared by Dr. Frances Swineford.

17.3 Other Models

If we assume a common value for the discriminating powers of the items, each $a_g = 1$, say, and take $D = 1$, we obtain the form

$$P_g(\theta) = \Psi(\theta - b_g) \equiv (1 + e^{b_g - \theta})^{-1}.$$

We can write

$$\theta^* = e^{\theta} \qquad \text{and} \qquad b_g^* = e^{b_g}$$

to denote, respectively, an ability parameter and an item difficulty parameter, each represented on a transformed scale. Then we have

$$P_g(\theta) \equiv P_g^*(\theta^*) = \left(1 + \frac{b_g^*}{\theta^*}\right)^{-1} = \frac{\theta^*}{b_g^*}\left(1 + \frac{\theta^*}{b_g^*}\right)^{-1}.$$

Rasch (1960) has developed the test model of this restricted logistic form. We see that this model is a special case of the logistic model in which all items have the same discriminating powers, and all items can vary only in their difficulties. Whenever this special logistic model holds, the considerable body of theoretical and practical methods developed by Rasch is applicable (see Chapter 21).

One very important question emerges at this point: Do the items in a test really differ from each other in discriminating power? This question is crucial to evaluating the validity of the models and methods of this and the following three chapters and to comparing these evaluations with evaluations of the validity of the simpler models and methods of Chapter 21. Some available item analysis data suggest an affirmative answer for multiple-choice paper and pencil tests. These data, which are represented in Table 17.3.1, are based on a sample of 3805 examinees. The table shows the mean and standard deviation of the sample biserial correlation between item score and test score for each of five different tests. If the true biserial correlation is 0.50 in a normal population of this size ($N = 3805$), then the standard error of a biserial correlation from a sample of this size will only be from about 0.016 to about 0.019, depending on the item difficulty. (An approximate formula appears in McNemar, 1962, Eq. 12.3.) Since the standard deviations in this particular sample are at least five times as large as this standard error, it is clear that the variation found here among item-test biserials is almost entirely due to real differences among the item discriminating power parameters. In this sample we find that even if we disregard the five percent of the items with the highest and the five percent with the lowest discriminating power parameters, we still have a range from about 0.31 to about 0.67. Since item-test biserials approximate item-ability biserials, whose close relation to the slope of the item characteristic curve was discussed rather fully in Section 16.10, it is clear that the item characteristic curves of the items in Table 17.3.1 differ from each other by more than a mere translation (change of origin).

If $\theta > b_g$, then for any fixed values of b_g and θ,

$$\Phi[a_g(\theta - b_g)] \qquad \text{and} \qquad \Psi[Da_g(\theta - b_g)]$$

both increase to 1 as a_g increases; and if $\theta < b_g$, then both decrease to 0 as a_g increases. We may represent these limiting values formally as

$$\Phi[\infty(\theta - b_g)] = \Psi[\infty(\theta - b_g)] \equiv \begin{cases} 1 & \text{if} \quad \theta > b_g, \\ 0 & \text{if} \quad \theta < b_g, \end{cases}$$

since

$$(\theta - b_g)\infty \equiv \begin{cases} +\infty & \text{if} \quad \theta > b_g, \\ -\infty & \text{if} \quad \theta < b_g. \end{cases}$$

For convenience, we can give the value 1 to the otherwise undefined symbols $\Phi(\infty \cdot 0)$, $\Psi(\infty \cdot 0)$. Then we may define an item characteristic curve by

$$P_g(\theta) = \Phi[\infty(\theta - b_g)] \qquad \text{or} \qquad \Psi[\infty(\theta - b_g)].$$

These may be considered extreme, limiting cases of ICCs within the normal ogive and the logistic test models. Such ICCs do not have the property, generally assumed above, of increasing continuously and strictly as θ increases. Each is characterized fully by a single difficulty parameter b_g; for abilities $\theta < b_g$ it has the value zero, and at this ability level it increases discontinuously to unity. These curves may be regarded as representing items whose responses y_g are error-free indicants of abilities, in the sense that taking $y_g = 1$ as indicating $\theta \geq b_g$ and $y_g = 0$ as indicating $\theta < b_g$ entails probability zero of erroneous indications for each possible value of θ. It may be said that ICCs of this extreme form give "perfect scaling", since an ordering of subjects' abilities θ on the basis of any test consisting of such items is error-free (with probability 1, or certainty). Such items are basic to the scaling methods and the theory developed by Guttman (1950), particularly in connection with scaling of latent traits θ representing attitudes.

Lazarsfeld has developed several classes of latent trait models, but primarily for the investigation of attitudes rather than abilities. One of these may conveniently be described here. If

$$P_g(\theta) = a_g(\theta - b_g), \qquad g = 1, \ldots, n,$$

where θ is restricted to an interval on which all values of $P_g(\theta)$ lie between 0 and 1, we have the *linear model.* Despite quantitative differences, here, as in other models described above, the item parameter a_g represents discriminating power in the sense of rate of change of $P_g(\theta)$ with respect to θ, and b_g locates the part of the θ scale where the item is effective. In Chapter 24, we shall present several other models developed by Lazarsfeld.

Methodological problems related to these models are discussed briefly by Torgerson (1958, Ch. 13), who gives references to basic papers and subsequent work. A later discussion is that of Lazarsfeld (1959).

Even subjects of very low ability will sometimes give correct responses to multiple-choice items, just by chance. One model for such items has been suggested by a highly schematized psychological hypothesis. This model assumes that if an examinee has ability θ, then the probability that he will *know* the correct answer is given by a normal ogive function $\Phi[a_g(\theta - b_g)]$ of exactly the kind considered in Section 16.5; it further assumes that if he does not know it he will guess, and, with probability c_g, will guess correctly. It follows from these assumptions that the probability of an incorrect response is

$$Q_g(\theta) = \{1 - \Phi[a_g(\theta - b_g)]\}(1 - c_g),$$

and that the probability of a correct response is the item characteristic curve

$$P_g(\theta) = c_g + (1 - c_g)\Phi[a_g(\theta - b_g)]. \tag{17.3.1}$$

The psychological hypothesis implicit here has been mentioned primarily to point up a mathematical feature of this form; the empirical validity of this form is not dependent on this psychological hypothesis. This model is possibly more reasonable than the random-guessing models discussed in Sections 14.3 and 14.5.

The function (17.3.1) approaches its minimum c_g as θ decreases. Its graph is that of a normal ogive curve except that the range of ordinates 0 to 1 is replaced by the range c_g to 1. If one of five multiple-choice alternatives were chosen at random whenever guessing occurred, we would have $c_g = \frac{1}{5}$, as in Fig. 17.3.1, where the other item parameters are equal to those in Fig. 16.5.1. Each of the general illustrative comments above concerning the item parameters a_g and b_g of normal ogive models can be adapted to apply to their roles in these models.

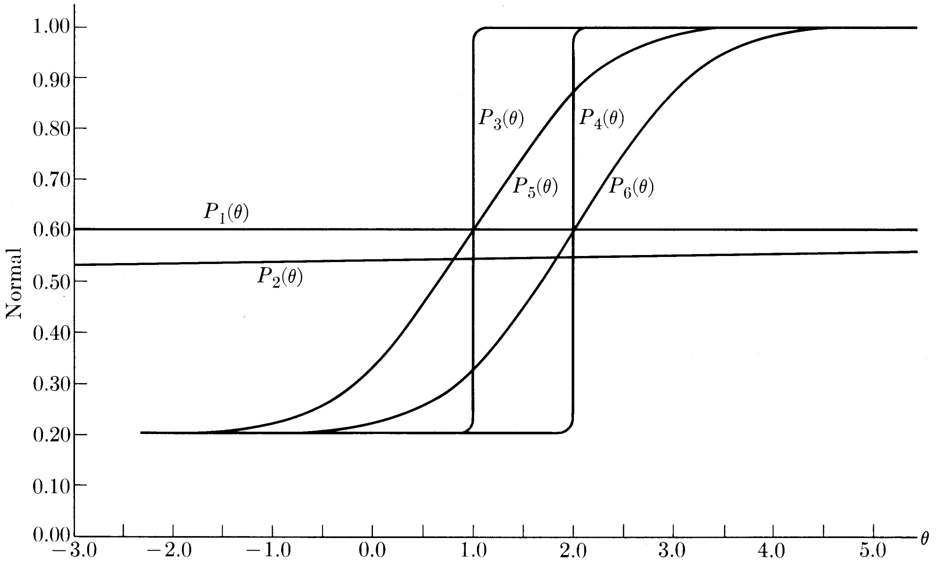

FIG. 17.3.1. Three-parameter normal ogive and logistic item characteristic curves.

Similarly with the logistic model, we may take account of guessing probabilities by using modified item characteristic curves, which here assume the form

$$P_g(\theta) = c_g + (1 - c_g)\Psi[Da_g(\theta - b_g)],$$

which Fig. 17.3.1 serves to illustrate. More detailed consideration of the roles of item parameters in such models is given below.

17.4 The Test as a Measuring Instrument: Examples of Classification and Estimation of Ability Levels by Use of Test Scores

We shall find it useful to consider the mathematical model of a test as having dual but related purposes. One purpose is to determine the value θ of an examinee's ability with adequate precision; the second is to classify an examinee into ability categories with adequately small probabilities of misclassification. We shall present brief descriptions of some estimation and classification methods based on test scores. These will illustrate some of the applications of the theory that we shall develop. Each of the simplifying assumptions or restrictions made here will require critical reconsideration later.

We shall consider a model of a test, represented by a specified probability function

$$\text{Prob}\,[\mathbf{V}' = (u_1, \ldots, u_n) \mid \theta],$$

possibly having one of the forms described above, in which the ability θ is the only unknown parameter. We shall adopt a specified test score formula $x = x(\mathbf{v}) \equiv x(u_1, \ldots, u_n)$. These two functions determine the cdf of the test score:

$$F(x \mid \theta) = \text{Prob}\,[X(\mathbf{V}) \leq x \mid \theta] \equiv \sum_{X(V) \leq x} \text{Prob}\,(\mathbf{V} = \mathbf{v} \mid \theta). \qquad (17.4.1)$$

Numerical determinations of $F(x \mid \theta)$ for a number of score formulas and tests will be illustrated. In the simplest case, that of items having identical characteristic curves

$$P \equiv P(\theta) \qquad \text{and} \qquad x = \sum_{g=1}^{n} u_g,$$

the cdf, $F(x \mid \theta)$, is just the binomial cdf for n trials with parameter $P(\theta)$:

$$F(x \mid \theta) = \sum_{k=0}^{x} \binom{n}{k} P^k Q^{n-k}, \qquad x = 0, 1, \ldots, n. \qquad (17.4.2)$$

Local independence is assumed here.

In most cases of interest, the magnitudes of discontinuities in $F(x \mid \theta)$ (that is, the probabilities of the individual possible values of x) will all be small for each θ, usually of the order of several percent or less. For many theoretical and practical purposes, it is convenient to treat $F(x \mid \theta)$ as continuous in x for each fixed θ, and also it is sometimes convenient to employ specific continuous functions of x as working approximations subject to appropriate bounds or independent checks on the approximations entailed. For illustrative simplicity, we treat $F(x \mid \theta)$ in this section as continuous and assume that for each fixed θ, it is strictly increasing from 0 to 1 with x. In the preceding binomial example, the convenient approximation is the usual one by the normal cdf (see, for example, Lindgren, 1962, p. 149):

$$F(x \mid \theta) \equiv \sum_{k=0}^{x} \binom{n}{k} P^k Q^{n-k} \doteq \Phi\left[\frac{x + \frac{1}{2} - nP}{(nPQ)^{1/2}}\right], \qquad (17.4.3)$$

which is continuous in x for each θ.

We further assume throughout this section that for each fixed value of x, $F(x \mid \theta)$ is strictly and continuously decreasing from 1 to 0 with θ; the respective distributions of x are said to be *stochastically ordered* when this condition holds. In the binomial example, this condition holds for both the exact and approximate formulas for $F(x \mid \theta)$, given that $x < k$. This condition is entailed by weak assumptions which are usually satisfied, namely, that each $P_g(\theta)$ increases strictly and continuously with θ, and that $x(u_1, \ldots, u_n)$ is nondecreasing in each u_g and increasing in at least one of them. The latter conditions hold in all cases described above.

When these conditions hold, the respective cdf's of scores of a given test can be represented conveniently in the manner illustrated in the schematic graphs of Figs. 17.4.1 and 17.4.2. Figure 17.4.2 is a schematic representation

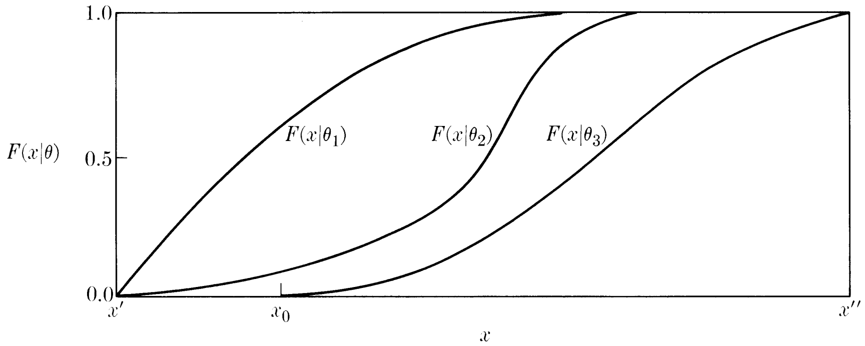

FIG. 17.4.1. Cdf's of scores x for several θ-values, $\theta_1 < \theta_2 < \theta_3$.

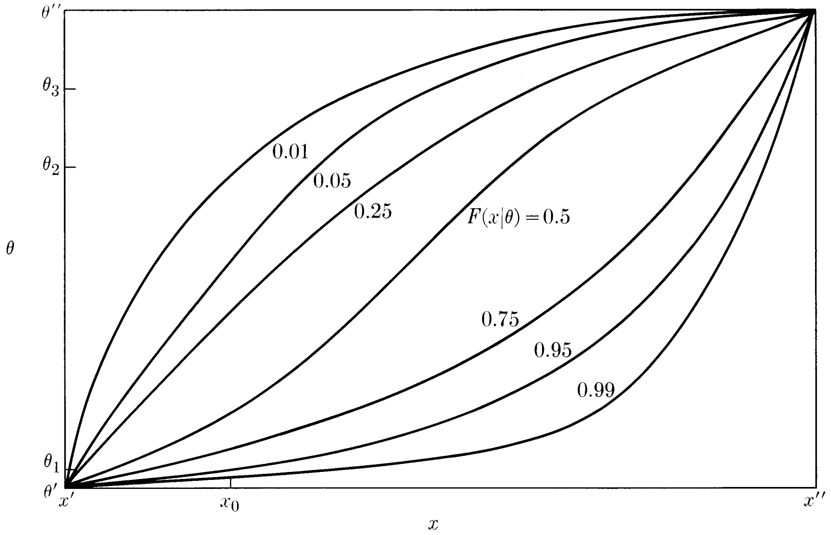

FIG. 17.4.2. Contours of constancy for cdf's $F(x, \theta)$.

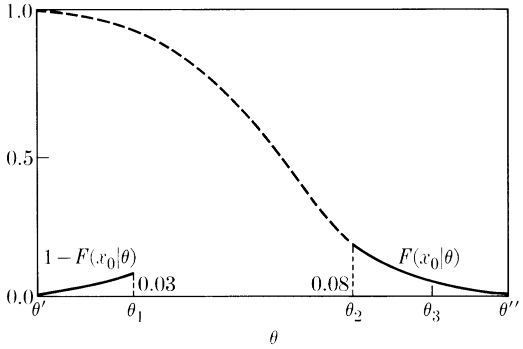

FIG. 17.4.3. Error probabilities of the classification rule that classifies high when $x > x_0$.

of the function of two arguments $F(x \mid \theta)$ that map several "contours of constant height" of the $F(x \mid \theta)$ "surface" over the (x, θ) plane. Figure 17.4.1 represents three "sections" ("slices") through this surface, made at $\theta = \theta_1, \theta_2, \theta_3$, respectively. Figure 17.4.3 represents in two forms one "section" made in the perpendicular direction at $x = x_0$; we shall explain this figure below.

The discriminating power of a test is illustrated most simply in problems of discriminating between just two levels of ability. One common rule classifies those subjects whose scores exceed some specified number x_0 as "high" and classifies others as "low". With this rule, if θ is any ability level considered definitely high, then $F(x_0 \mid \theta) = \text{Prob}\,(X \leq x_0 \mid \theta)$ is the probability of erroneous (low) classification of a subject of that ability. Since $F(x_0 \mid \theta)$ decreases as θ increases, it is natural to focus attention on the smallest θ value considered definitely high, say θ_2. The rule's maximum probability of erroneous classification of a high-ability subject is then $F(x_0 \mid \theta_2)$, as illustrated by Fig. 17.4.3. Similarly, if θ_1 is the highest ability considered definitely low, then

$$1 - F(x_0 \mid \theta_1) = \text{Prob}\,(X > x_0 \mid \theta)$$

is the rule's maximum probability of erroneous classification of a low-ability subject. At abilities between θ_1 and θ_2, neither classification is considered definitely erroneous and no error probabilities are considered.

By decreasing x_0, we can decrease $F(x_0 \mid \theta_1)$, the maximum misclassification probability for low abilities, but only at the cost of increasing $1 - F(x_0 \mid \theta_2)$, the maximum misclassification probability for high abilities. Evidently the possibility of circumventing such restrictions on the discriminating power attainable with a given test depends on basic reconsideration of the forms of test-score formulas and classification rules adopted; and these considerations might show that improvement requires the use of a different test.

We note that some of the present considerations parallel some of the interpretations given above of the discriminating power of single items in terms of item characteristic curves. The common element is the role of the rate of increase of $P_g(\theta)$ and of $1 - F(x_0 \mid \theta)$, respectively, as θ increases. So long as a test is used only to provide a classification rule based on a comparison of its scores with some fixed critical value x_0, the test is in effect equivalent to a single hypothetical test item having responses

$$u_1^* = \begin{cases} 1, & \text{corresponding to} \quad x > x_0 \\ 0, & \text{corresponding to} \quad x \leq x_0, \end{cases}$$

and item characteristic curve

$$P_1^*(\theta) = 1 - F(x_0 \mid \theta).$$

We can consider parameters describing the form of $F(x_0 \mid \theta)$ and $1 - F(x_0 \mid \theta)$ in rough analogy with the parameters of single items: For example, if $F(x_0 \mid \theta') = \frac{1}{2}$, then θ' can be called the *difficulty level of the classification rule;* and

$$-\frac{\partial}{\partial\theta} F(x_0 \mid \theta)$$

evaluated at $\theta = \theta'$ can be called the *discriminating power of the classification rule.* We shall consider in detail below the ways in which such parameters of the test and other properties of classification rules depend on the parameters of the respective test items. Parameters such as a_g, b_g in logistic or normal items serve to characterize an item fully, and these parameters admit heuristically useful and relevant descriptive interpretations. However, their principal significance lies in their precise role in contributing to the information structure of a test, a notion we shall elaborate in the following sections and chapters. For a classification rule represented by a function $1 - F(x_0 \mid \theta)$, an analogous pair of parameters may be of some limited descriptive value, but in general they must fall far short of determining fully the course of $1 - F(x_0 \mid \theta)$ and the values of all error probabilities of practical interest. A summary description of the error probabilities that is more useful for many purposes is a pair of points such as those represented in Fig. 17.4.3, which indicate that at the values θ_1, θ_2 the error probabilities of respective types are 0.03 and 0.08.

A standard technique of estimation, that of confidence limits, is directly applicable when the distributions of test scores are available graphically, as in Fig. 17.4.3, or equivalently in tables of percentage points. A lower confidence limit estimator with a confidence coefficient of 95%, say, is defined as any statistic $t(\mathbf{v})$ having the property that

$$\text{Prob}\,[t(\mathbf{V}) \leq \theta \mid \theta] = 0.95 \qquad \text{for each } \theta.$$

That is, for each possible value θ of an examinee's ability, the probability is 0.95 that the estimate $t(\mathbf{v})$ derived from the response pattern of such an examinee will be a correct lower bound on his ability.

In the case at hand, where $\mathbf{v}$ is represented just by a test score x, it is easy to obtain a statistic $t(x)$ with the above property. Let x^* denote the numerical test score of an examinee. Let $\theta^*(x^*, 0.95)$ denote the number θ^* that satisfies the equation $F(x^* \mid \theta^*) = 0.95$; in Fig. 17.4.2, θ^* corresponds to x^* in the sense that (x^*, θ^*) is a point on the 0.95 contour. Then θ^* is a lower 95% confidence limit estimate of the examinee's ability θ. (The fact that Prob $[\theta^*(X, 0.95) \leq \theta \mid \theta] = 0.95$ is an easily derived consequence of the definition of θ^*.) Taking $\theta^* = 1.3$, for concreteness of illustration, we may record this conveniently in the notation: Conf $(\theta \geq 1.3) = 0.95$.

Other confidence limits are determined similarly. For example, $\theta^*(X, 0.25)$ is an upper 75% confidence limit estimator, defined implicitly by $F(x \mid \theta) = 0.25$ and having the basic property that

$$\text{Prob}\,[\theta^*(X, 0.25) > \theta \mid \theta] = 0.75 \qquad \text{for each } \theta.$$

The pair of estimators, $\theta^*(x, 0.95)$, $\theta^*(x, 0.05)$, together constitute a 90%

confidence interval estimator of θ; For each possible true value θ, they include θ between them with probability 90%. Among the various types of useful point estimators of θ, one which we may conveniently describe here is $\theta^*(x, 0.5)$. This point estimator is median-unbiased, that is, it both overestimates and underestimates θ with probability $\frac{1}{2}$.

The precision of a confidence interval estimator is represented by its confidence coefficient, together with the typical lengths of the interval estimates that it determines; or, more precisely and adequately, by error probabilities for over- or underestimation by various amounts. We shall indicate below how such precision properties of confidence intervals and confidence limits can be related in detail to the discriminatory power of a test in classification by ability levels.

17.5 The Information Structure of a Test and Transformations of Scale of Scores

When we apply a test model in conjunction with a specific test to such classification and estimation problems as the ones illustrated in the preceding section, we observe that no properties of the model play any role except the cdf's of the score of that specific test. For example, Fig. 17.4.2 might represent two different tests, each with very different numbers and types of items and item characteristic curves, but the estimation and classification methods based on scores of the respective tests would still have identical error-probability properties. This equivalence would hold even if the cdf's were different, but could be made to coincide when scores x of one test were transformed by a suitable increasing function $x^*(x)$ into scores x^* of the second test. This is true because no properties of the scale of scores x beyond simple ordering have been used here. Thus, for such standard inference methods based on an adopted test score formula, we may consider the family of distributions of scores $F(x \mid \theta)$ as representing the essential *information structure* or *canonical form* of a test, with the qualification that the scale of scores x plays only the role of simple ordering.

To illustrate this qualification, consider any given family of cdf's $F(x \mid \theta)$, any arbitrarily chosen ability θ_2, and the function defined by

$$x^* \equiv x^*(x) = F(x \mid \theta_2).$$

This is a strictly increasing function of x, and we can adopt it to define scores x^* on a new scale; the range of such scores is $0 \leq x^* \leq 1$. Let the cdf's of such scores x^* be denoted by

$$F^*(x^* \mid \theta) \equiv \text{Prob}\,[X^*(\mathbf{V}) \leq x^* \mid \theta], \qquad 0 \leq x^* \leq 1.$$

A special property of scores defined in this way is that when $\theta = \theta_2$ [that is, the ability level that has been arbitrarily chosen for the definition of $x^*(x)$], the distribution of scores X^* takes the special "uniform" form

$$F^*(x^* \mid \theta_2) \equiv x^*, \qquad 0 \leq x^* \leq 1.$$

This property characterizes the *probability integral transformation* $x^*(x)$. An illustration appears in Fig. 17.5.1, a figure that is a transformed version of Fig. 17.4.1. Since such transformations of scores are typically nonlinear, expected values and variances of the transformed scores x^* do not have any simple relations to expected values and variances of scores x on the original scale. Thus the concepts and methods presented in this section are not closely linked with any use of moments of distributions of test scores at given ability levels.

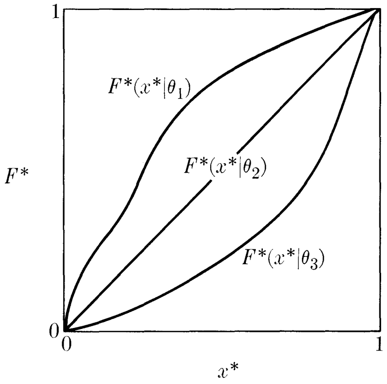

FIG. 17.5.1. Transformed version of Fig. 17.4.1.

[If the curve for θ_3 were deleted from Fig. 17.5.1 and the resulting figure were rotated, the new figure would be familiar to many students of mathematical statistics (see, for example, Lindgren, 1962, p. 236). For each α, the test of the hypothesis H_0: $\theta = \theta_2$ against the hypothesis H_1: $\theta = \theta_1$, based on rejecting H_0 just when x^* is sufficiently small, is the test that rejects just when $x^* \leq x_\alpha^* \equiv \alpha$. The power of this test is given by $1 - \beta = F^*(x_\alpha^* \mid \theta_1) \equiv F^*(\alpha \mid \theta_1)$. Thus $\beta = 1 - F^*(\alpha \mid \theta_1)$ gives the "α, β curve".]

17.6 Transformations of Scales of Ability

It is interesting to consider the preceding point concerning the scaling of *scores,* in combination with the point concerning the scaling of *abilities* illustrated at the end of Section 17.1, where a certain freedom in specification of the ability scale was discussed. The latter point can be applied here: Abilities θ can be replaced by abilities $\theta^* = \theta^*(\theta)$ on a transformed scale in such a way that the family of cdf's of scores

$$F^{**}(x^* \mid \theta^*) = \text{Prob}\,[X^*(\mathbf{V}) \leq x^* \mid \theta^*]$$

is given any chosen form compatible with the other conditions thus far assumed. For example, the transformation $\theta^*(\theta)$ defining the new scale of abilities can be chosen so that each possible score value x^* is the median of the distribution of scores for ability level $\theta^* = \theta^*(\theta) = x^*$; that is, so that

$$F^{**}(x^* \mid \theta^*) = \tfrac{1}{2} \qquad \text{whenever} \qquad x^* = \theta^* = \theta^*(\theta),$$

as in Fig. 17.6.1, which is a transformed version of Fig. 17.4.2. To prove this, we note that the condition $F(x \mid \theta) = 0.5$ defines implicitly the function $x(\theta)$, the median of scores x for each ability θ, in terms of the given cdf's. Hence $x^*[x(\theta)]$ is the median of transformed scores x^* for each ability θ. We are now free to define a transformation of abilities by

$$\theta^*(\theta) = x^*[x(\theta)].$$

Now for each ability θ, the transformed ability $\theta^* = \theta^*(\theta)$ coincides with the median of the distribution of transformed scores.

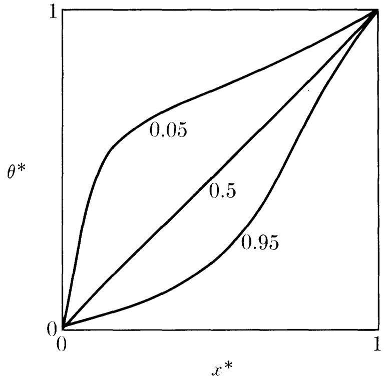

FIG. 17.6.1. Transformed version of Fig. 17.4.2.

We may mention another significant example of possible rescaling of scores and abilities. Let $x^*(x)$ be any arbitrarily chosen strictly increasing function, subject only to the mild restriction that the expected values of scores, $\mathscr{E}(X^* \mid \theta)$, exist for each θ. Let us determine a new scale of scores by the transformation $x^*(x)$. Next, we can choose a transformed scale of abilities θ^*, determined by the transformation function $\theta^*(\theta) = \mathscr{E}(X^* \mid \theta)$. From the assumption that the cdf's $F(x \mid \theta)$ are stochastically ordered, it follows that $\theta^*(\theta)$ is an increasing function, and that the cdf's $F^{**}(x^* \mid \theta^*)$ will also be stochastically ordered. Every scale of abilities θ^* that may be determined in this way satisfies the essential condition for the definition of true score presented in Chapter 2, namely, $\mathscr{E}(X^* \mid \theta^*) = \theta^*$ for each θ^* (see Chapter 24).

It is interesting to consider an analogous question: If any test model and score formula are given and are represented by a specified family of cdf's $F(x \mid \theta)$ having the two monotonicity properties assumed above, then is it always possible to keep the *given ability scaling* (which, of course, may have been obtained by an arbitrary transformation from a previous ability scaling), and also to realize simultaneously, by means of some monotone transformation $x^*(x)$ of the score scale, the essential condition for true-score theory, namely, $\theta = \mathscr{E}(X^* \mid \theta)$ for each θ? The answer is, usually, no—the possibility depends on the detailed structure of the given cdf's $F(x \mid \theta)$. To illustrate this simply, we shall assume that x has a finite number of possible values

$$x_1 < \cdots < x_j < \cdots < x_M,$$

and consider an arbitrary sequence of different possible values of θ, namely, $\theta_1, \theta_2, \ldots \theta_i, \ldots$. Let $C_{ij} = \text{Prob}\,(X = x_j \mid \theta_i)$ for each i, j. (Here we drop the assumption of continuity of the cdf's $F(x \mid \theta)$, an assumption which is typically inexact although useful elsewhere.) If $x^*(x)$ is any monotone transformation, we may write

$$x_j^* = x^*(x_j) \qquad \text{and} \qquad x_1^* < \cdots < x_j^* < \cdots x_M^*.$$

If the transformed scores x^* are to satisfy the true score assumption

$$\theta = \mathscr{E}(X^* \mid \theta) \equiv \mathscr{E}[x^*(X) \mid \theta], \quad \text{for each } \theta,$$

then for each i we must have

$$\theta_i = \mathscr{E}[x^*(X) \mid \theta_i] \qquad \text{or} \qquad \theta_i = \sum_{j=1}^{M} C_{ij} x_j^*.$$

In general, such linear equations in M unknowns x_j^* are inconsistent, even when only $M + 1$ such equations (determined by any chosen $M + 1$ values θ_i) are considered in isolation. Thus the possibility of realizing the conditions for true score theory *for the given ability scale*, even by monotone transformation of the given score formula, is limited by and dependent on the detailed structure of the given model $F(x \mid \theta)$. This contrasts with the possibility of realizing the true score assumptions *for the given score scale*, which, as we have seen above, is always possible if a monotone transformation of the ability scale is allowed. The discussion here amplifies and formalizes the discussion in Chapter 2 of the relationship among various concepts of true score.

On the other hand, to explore the approximate applicability of classical true-score theory to a given model when the given ability scaling is to be retained, we can first choose successively values θ_i that seem to represent the range of abilities of interest effectively and can then consider the sequence of equations

$$\theta_i = \sum_{j=1}^{M} C_{ij} x_j^*,$$

continuing so long as the equations are consistent and allow ordered solutions x_j^*. If any set of such equations does not determine unique, ordered solutions, we may supplement it by adding arbitrary and possibly convenient independent linear restrictions on the x_j^*, possibly including specification of convenient values for

$$x_1^*, \quad x_M^*, \quad \frac{x_1^* + x_M^*}{2}, \quad \text{or} \quad \frac{1}{M}\sum_{j=1}^{M} x_j^*,$$

or some combination of these, until we have obtained M linearly independent equations.

Whenever $F(x \mid \theta)$ is a normal cdf for each θ, we may take $\theta^*(\theta) = \mathscr{E}(X \mid \theta)$, which is both the mean and the median of X, for each θ. When $F(x \mid \theta)$ is at least approximately a normal cdf, then $\theta^*(\theta) = \mathscr{E}(X \mid \theta)$ is usually approximately the median (as well as exactly the mean) of X.

Of course, weak true-score theory is characterized by its use of no restrictive assumptions on the forms of the cdf's $F(x \mid \theta)$ of scores other than low-order moments of scores. The preceding considerations illustrate some of the many connections and differences to be found between weak and strong true-score theories.

17.7 Calculations of Distributions of Test Scores

Applications of the inference methods illustrated above require adequate numerical determinations of the distributions of test scores at respective ability levels. In most practical work with cognitive tests, response patterns are represented only by test scores having the particular form

$$x = x(\mathbf{v}) = \sum_{g=1}^{n} w_g u_g \tag{17.7.1}$$

of weighted sums of item responses, where the w_g are specified numerical weights. Most commonly, the weights are specified as equal, either as $w_g \equiv 1$, where calculation of x then gives the number of correct responses, or as $w_g \equiv 1/n$, where calculation of x then gives the proportion of correct responses. In the following chapters, we shall see that in important cases a suitably chosen linear (or weighted-sum) score formula can be used to provide estimators with optimal or nearly optimal precision and classification rules of good discriminating power. In this section, we shall present some useful theoretical and computational methods for calculating distributions of test scores of this form and illustrate these by numerical examples of the applications illustrated above.

The principal result that we shall present here is the normal approximation to the cdf $F(x \mid \theta)$ for score formulas x of any weighted sum form $x = \sum_{g=1}^{n} w_g u_g$, where the w_g are given constants. The theoretical basis for such normal approximations in the general case consists of the central limit theorems available for sums of nonidentical independent random variables (see, for example, Lindgren, 1962, p. 147, and Loève, 1955, p. 288). The resulting approximation formulas for $F(x \mid \theta)$ depend on the given test model only through the mean and variance of X for each θ:

$$\mathscr{E}(X \mid \theta) = \sum_{g=1}^{n} \mathscr{E}(w_g U_g \mid \theta) \equiv \sum_{g=1}^{n} w_g P_g(\theta), \tag{17.7.2}$$

$$\sigma^2 (X \mid \theta) = \sum_{g=1}^{n} \sigma^2 (w_g U_g \mid \theta) \equiv \sum_{g=1}^{n} w_g^2 P_g(\theta) Q_g(\theta). \tag{17.7.3}$$

Then the approximation formula is

$$F(x \mid \theta) \doteq \Phi\{[x - \mathscr{E}(X \mid \theta)]/\sigma(X \mid \theta)\}. \tag{17.7.4}$$

In connection with various specific test models and problems of application below, the preceding general formulas for the moments of scores will be specialized and substituted in the last relation.

For test models with nonequivalent items, and for composite scores with unequal weights, we require here a form of the central limit theorem that allows nonidentically distributed terms $w_g U_g$. On the other hand, for many practical purposes we may conveniently interpret the hypothetical concept of increase without limit of the number n of nonequivalent test items as the case of a test model with $G_n = ng$ items specified as follows: The first n items may have any specified ICCs; each successive set may consist of n items equivalent, respectively, to those of the first set; and G may increase without limit. The simplest case of the central limit theorem, that of identically distributed terms, applies here, since each set of n items can formally be considered to contribute a single term

$$Z_r = \sum_{g=1}^{n} w_g U_{nr+g}, \quad r = 0, 1, 2, \ldots, \qquad \text{to} \qquad X = \sum_{r=0}^{G} Z_r,$$

provided that $w_{nr+g} = w_g$ for $r = 1, 2, \ldots$.

Examples: Moments and quantiles of test scores for items of various types.

Moments of item responses

$$\mathscr{E}(U_g \mid \theta) = P_g(\theta), \qquad \sigma^2(U_g \mid \theta) = P_g(\theta) Q_g(\theta),$$

1. Normal ogive

$$\mathscr{E}(U_g \mid \theta) = \Phi[L_g(\theta)], \qquad \sigma^2(U_g \mid \theta) = \Phi[L_g(\theta)]\Phi[\,-\,L_g(\theta)],$$

where $L_g(\theta) = a_g\theta - b_g$.

2. Logistic

$$\mathscr{E}(U_g \mid \theta) = \Psi[DL_g(\theta)], \qquad \sigma^2(U_g \mid \theta) = \psi[DL_g(\theta)],$$

where

$$\psi(t) = \frac{\partial}{\partial t}\Psi(t) \equiv \frac{e^t}{(1 + e^t)^2}.$$

3. Three-parameter logistic

$$\mathscr{E}(U_g \mid \theta) = c_g + (1 - c_g)\Psi[DL_g(\theta)] = \Psi[DL_g(\theta)] + c_g\Psi[-DL_g(\theta)],$$
$$\sigma^2(U_g \mid \theta) = (1 - c_g)\psi[DL_g(\theta)] + c_g(1 - c_g)\Psi[-DL_g(\theta)]^2.$$

Moments of terms in locally best composite scores (developed below in Section 19.3)

$$w_g(\theta) = P_g'(\theta)/P_g(\theta)Q_g(\theta),$$
$$\mathscr{E}[w_g(\theta)U_g \mid \theta] = [P_g'(\theta)/P_g(\theta)Q_g(\theta)]P_g(\theta) = P_g'(\theta)/Q_g(\theta),$$
$$\sigma^2[w_g(\theta)U_g \mid \theta] = w_g(\theta)^2\sigma^2(U_g \mid \theta) = [P_g'(\theta)^2/P_g(\theta)^2Q_g(\theta)^2]P_g(\theta)Q_g(\theta)$$
$$= P_g'(\theta)^2/P_g(\theta)Q_g(\theta).$$

1. Normal ogive

$$w_g(\theta) = a_g\varphi[L_g(\theta)]/\Phi[L_g(\theta)]\Phi[-L_g(\theta)],$$
$$\mathscr{E}[w_g(\theta)U_g \mid \theta] = \varphi[L_g(\theta)]/\Phi[-L_g(\theta)],$$
$$\sigma^2[w_g(\theta)U_g \mid \theta] = a_g^2\varphi[L_g(\theta)]^2/\Phi[L_g(\theta)]\Phi[-L_g(\theta)].$$

2. Logistic

$$w_g(\theta) = Da_g \quad \text{(uniformly best weights)},$$
$$\mathscr{E}(w_gU_g \mid \theta) = Da_g\psi[DL_g(\theta)], \qquad \sigma^2(w_gU_g \mid \theta) = D^2a_g^2\psi[DL_g(\theta)].$$

3. Three-parameter logistic

$$w_g(\theta) = Da_g\Psi[DL_g(\theta) - \log c_g], \qquad \mathscr{E}[w_g(\theta)U_g \mid \theta] = Da_g\Psi[DL_g(\theta)],$$
$$\operatorname{Var}[w_g(\theta)U_g \mid \theta] = (1 - c_g)D^2a_g^2\psi[DL_g(\theta) - \log c_g]\psi[DL_g(\theta)].$$

Moments of composite scores. Dividing each weight w_g in a scoring formula by the same positive constant (for example, the sum of the weights) does not change the ratio between respective weights, which is the essential feature of the scoring formula. Therefore we may express any composite score formula in the form

$$x = \frac{\sum_{g=1}^{n} w_g u_g}{\sum_{g=1}^{n} w_g}.$$

For example, the weights w_g may be

1) equal weights, for instance, $w_g = 1$ for $g = 1, \ldots, k$;

2) best weights (developed below in Section 18.4)

$$w_g(\theta_1, \theta_2) = \log \frac{P_g(\theta_2)Q_g(\theta_1)}{P_g(\theta_1)Q_g(\theta_2)};$$

or

3) locally best weights $w_g(\theta) = P_g'(\theta)/P_g(\theta)Q_g(\theta)$ as developed in Section 19.3. Thus we may write the moments of any composite score x as, say,

$$\mathscr{E}(X \mid \theta) = \frac{\sum_{g=1}^{n} w_g P_g(\theta)}{\sum_{g=1}^{n} w_g} \equiv \mu(\theta) \qquad \text{and} \qquad \sigma^2(X \mid \theta) = \frac{\sum_{g=1}^{n} w_g^2 P_g(\theta) Q_g(\theta)}{\left(\sum_{g=1}^{n} w_g\right)^2} \equiv \sigma^2(\theta).$$

Quantiles of composite scores under the normal approximation: a measure of information. Using our previous assumption that a given composite score x has cdf's $F(x \mid \theta)$ that are continuously strictly increasing in x and decreasing in θ, we may implicitly define the $(1 - \alpha)$-quantile of X, which we denote by

$$x^*(1 - \alpha, \theta),$$

as the solution x of

$$F(x \mid \theta) = 1 - \alpha. \tag{17.7.5}$$

If we assume in particular a normal form for the cdf's $F(x \mid \theta)$, we have

$$x^*(1 - \alpha, \theta)$$

defined as the solution x of

$$F(x \mid \theta) \equiv \Phi\{[x - \mu(\theta)]/\sigma(\theta)\} = 1 - \alpha. \tag{17.7.6}$$

Taking Φ^{-1} of both sides and solving for x, we then have

$$x^*(1 - \alpha, \theta) = \mu(\theta) + \Phi^{-1}(1 - \alpha)\sigma(\theta). \tag{17.7.7}$$

[The quantity $\Phi^{-1}(1 - \alpha)$ is a normal deviate cutting off a normal-curve left-tail area of $1 - \alpha$.]

The composite score will actually approach a normal form with increasing n, under the slight restriction that the values $w_g^2 P_g(\theta) Q_g(\theta)$ are uniformly bounded away from zero for the given θ-value considered. (This follows from the central limit theorem for the case of nonidentically distributed terms; see, for example, Loève, 1955, p. 310.) Under mild additional conditions (which will often be satisfied, and which can be checked with reference to specific applications), formula (17.7.7) can be approximated adequately closely, over any interval of θ-values centered at any given value θ' and appreciably wide, by a linear function of θ. This function of θ may be written

$$x^*(1 - \alpha, \theta) = A + B(\theta - \theta'), \tag{17.7.8}$$

where

$$A = \mu(\theta') + \Phi^{-1}(1 - \alpha)\sigma(\theta') \qquad \text{and} \qquad B = \mu'(\theta'), \qquad \mu' = \frac{\partial}{\partial \theta}\mu(\theta).$$

[This function represents the Taylor series approximation to (17.7.7),

$$\begin{aligned} x^*(1-\alpha,\theta) \doteq \mu(\theta') + \Phi^{-1}(1-\alpha)\sigma(\theta') + \mu'(\theta')(\theta-\theta') \\ + \Phi^{-1}(1-\alpha)\sigma'(\theta')(\theta-\theta'), \end{aligned}$$

further simplified by deleting the last term. This term may be deleted because

$$\sigma'(\theta') = \frac{\partial}{\partial\theta}\sigma(\theta)|_{\theta=\theta'}$$

tends to be negligible in comparison with $\mu'(\theta')$.] By solving (17.7.8) for θ, we obtain the corresponding linear approximation

$$\theta^*(x, 1-\alpha) = \theta' + [x - \mu(\theta') - \Phi^{-1}(1-\alpha)\sigma(\theta')]/\mu'(\theta'). \tag{17.7.9}$$

Now the latter formula represents (approximately) the lower $(1-\alpha)$-level confidence limit estimate of the ability θ of an individual with score x, as discussed in Section 17.4 above. One natural and convenient indication of the value of a given test and scoring formula is the width of the resulting confidence interval estimates of ability. The width of the approximate $(1-2\alpha)$-level confidence interval indicated by the approximation (17.7.9) for any given $\alpha < \frac{1}{2}$ is just $\theta^*(\alpha, x) - \theta^*(1-\alpha, x)$ as determined from (17.7.9):

$$[\Phi^{-1}(1-\alpha) - \Phi^{-1}(\alpha)]\frac{\sigma(\theta')}{\mu'(\theta')}\cdot$$

We see that this width is proportional to $\sigma(\theta')/\mu'(\theta')$, a constant independent of α, x, and θ under the assumed approximation. For θ near θ', therefore, this constant serves as an index of precision of interval estimation based on the given test and scoring formula. As it turns out, the same constant also characterizes the effectiveness of the test and scoring formula for a wide variety of other purposes. Hence we shall use the term *information* to designate the related quantity

$$I(\theta', x) = \mu'(\theta')^2/\sigma^2(\theta'). \tag{17.7.10}$$

More precisely, we shall refer to $I(\theta', x)$ as the *information* provided by the given test and composite scoring formula in the neighborhood of θ'. The function $I(\theta, x)$ is called *the information function of the scoring formula x*. It should be noted that the symbol x appears here not as a variable argument of $I(\theta, x)$, but as an abbreviation for "the probability distributions $F(x, \theta)$ of the scoring formula x", in terms of which $I(\theta, x)$ is defined. The definition is of course made with reference to some specified mental test, in terms of which the scoring formula is defined; thus, for a given scoring formula, $I(\theta, x)$ is a function of θ only.

There are two additional reasons for using the term "information" for this quantity:

1. Let us consider the error probability functions of classification rules based on x. At θ', the slope of each of these functions has a given value $\alpha = \alpha(\theta')$. With increasing n, these values tend to be proportional to

$$\sqrt{I(\theta', x)} \equiv \mu'(\theta')/\sigma(\theta').$$

Proof. Writing $F(x \mid \theta) = \Phi\{[x - \mu(\theta)]/\sigma(\theta)\} = \Phi(t)$, where $t = [x - \mu(\theta)]/\sigma(\theta)$, we see that the slope of the error probability function $1 - F(x \mid \theta)$ is $-\partial F(x \mid \theta)/\partial\theta$. By the chain rule, this can be written as $-[d\Phi(t)/dt]\,(\partial t/\partial\theta)$, or $-\varphi(t)\,(\partial t/\partial\theta)$. Hence

$$\begin{aligned}\frac{\partial t}{\partial\theta} &= -\frac{\{\sigma(\theta)\mu'(\theta) + [x - \mu(\theta)]\sigma'(\theta)\}}{\sigma^2(\theta)} = -\frac{\mu'(\theta)}{\sigma(\theta)} - t\frac{\sigma'(\theta)}{\sigma(\theta)}\\ &\doteq -\sqrt{I(\theta, x)},\end{aligned}$$

since, as we noted in the derivation of (17.7.9), $\sigma'(\theta)$ is small compared with $\mu'(\theta)$.

2. With increasing n, when θ' is the true value, the point estimator $\theta^*(x, 0.5)$ tends to be normally distributed, with mean θ' and variance $1/I(\theta', x)$.

Thus $I(\theta, x)$ plays the role of an index of precision of estimation. If we are dealing with nonlinear scoring functions $x = x(\mathbf{v})$, then we cannot apply the central limit theorem in the direct way indicated in connection with (17.7.4) above. Nevertheless, for an important and wide class of nonlinear scoring functions and estimators, we can show that there is an approach to a limiting normal distribution with increasing n. The definition $I(\theta, x)$ or $I(\theta, \theta^*) = \mu'(\theta)^2/\sigma^2(\theta)$ is extended to such cases of nonlinear x or θ^* by taking $\mu(\theta)$ and $\sigma^2(\theta)$ to represent the *asymptotic moments* of x and θ^*. These asymptotic moments are moments of the limiting normal distributions, which are in theory, and in relevant examples, distinct from the limits of exact moments of x or θ^*.

In particular, in Section 20.3, we shall consider the maximum likelihood estimator $\hat{\theta}$ and its information function $I(\theta, \hat{\theta})$ in some detail. As in the preceding special case, we shall see that the role of an index of precision of estimation is played quite frequently by the information function $I(\theta, x)$, for a given scoring formula, and $I(\theta, \theta^*)$, for a given test and estimator.*

* For derivations and discussions of these properties, see Cramér (1946, pp. 498–506) or Birnbaum (1961a, pp. 122–127). In such discussions of asymptotic distributions in connection with maximum likelihood, the results (1) and (2) above are obtained by replacing the "score" $S(x, \theta) = (\partial/\partial\theta) \log f(x \mid \theta)$ by $[x - \mathscr{E}(X \mid \theta)]/\sigma(X \mid \theta)$.

These and other uses and interpretations of the information functions $I(\theta, x)$ of various test models and composite score formulas will appear below, particularly in Chapter 20, where self-contained discussions of some aspects of information functions are given.

17.8 Quantal Response Models in General

The test models introduced in this chapter have analogues in other technical and scientific areas. Models of the general form Prob $(\mathbf{V} = \mathbf{v} \mid \theta)$ have been called *quantal response models.* The normal ogive model (including the three-parameter case described above) has been used extensively in biological assay work. (See, for example, Finney, 1944 and 1952. In the second reference, comparisons between biological and psychometric applications are given.) The use of the logistic model as an alternative to the normal in bioassay work has also been developed extensively (see Berkson, 1953 and 1957). For another type of biological assay, the *dilution series* model with $P_g(\theta) = 1 - e^{-a_g\theta}$ has been used (Fisher, 1922, pp. 363–366, and Cochran, 1950). Applications of such models have also been made in industrial gauging (Stevens, 1948) and genetics (for example, Rao, 1965, pp. 302–309, and Kempthorne, 1957, p. 181, and references therein). An appreciable part of the discussion in the next chapters has general relevance to quantal response models.

17.9 Estimation of Item Parameters

Two maximum likelihood methods have been given for estimating the item parameters in the normal ogive test model, by Tucker (1951) and by Lord (1953). These are discussed by Torgersen (1958, pp. 388–391), where they are related to other mathematical problems that arise in scaling. In the following paragraphs (1) and (2), we present two adaptations of these methods to the case of the logistic model. [For the restricted case of the logistic model described in Section 17.2, in which only the item difficulty parameters b_g are unknown, Rasch (1960) has given advantageous estimation methods. Many details of the derivation and calculation of estimates presented in the next paragraphs have forms similar to those of the more restricted estimation problem discussed in more detail in Section 20.3, which deals with maximum likelihood estimates of ability.]

The likelihood function of the responses observed when an n-item test is administered to a group of N examinees of abilities $\theta_1, \theta_2, \ldots, \theta_N$ is

$$L = \prod_{c=1}^{N} \prod_{g=1}^{n} \{1 - \Psi[Da_g(\theta_c - b_g)]\} \exp[Da_g(\theta_c - b_g)u_{gc}]. \qquad (17.9.1)$$

Let

$$x_c = \sum_{g=1}^{n} u_{gc}$$

denote the raw score of examinee c. Then

$$\mathscr{E}(X_c \mid \theta_c) = \sum_{g=1}^{n} \Psi[Da_g(\theta_c - b_g)]$$

is an increasing function of θ_c, provided that all a_g are positive. For two examinees of abilities θ_c and $\theta_{c'} > \theta_c$, we have Prob $\{X_{c'} > X_c\} \to 1$ as n increases, provided that the a_g are bounded away from zero and the b_g are bounded. That is, there is a tendency for ability order to be reflected correctly in the ordering of raw scores, as the number of items increases.

1. If we assume that the examinees are a random sample from a population in which the ability θ has a standard normal (or logistic) distribution, then, as N increases, the distribution of θ_c values over examinees converges (with probability one) to the standard normal (or logistic) distribution. Correspondingly the ability $\theta_{[PN]}$, which exceeds just a given proportion P of the abilities θ_c in a sample of n examinees, converges (with probability one), as n increases, to $\theta_P = \Phi^{-1}(P)$ [or to $\Psi^{-1}(P)/D$]. This second limit is the ability that exceeds just the proportion P of abilities in the population. Let P_c denote the proportion of raw scores in the sample that are less than x_c, and let

$$\theta(x_c) = \Psi^{-1}(P_c)/D. \tag{17.9.2}$$

Then it follows, under the conditions on item parameters mentioned in the preceding paragraph, that $\theta(x_c) \to \theta_c$ (with probability one) as both n and N increase. Thus, in practice, with N and n finite, we may regard $\theta(x_c)$ as an estimate of θ_c. In the next paragraphs, we treat the θ_c as known, with the understanding that in applications they shall be replaced by their numerical estimates $\theta(x_c)$.

The likelihood function L now has as unknown arguments just the $2n$ item parameters a_g and b_g. The maximum likelihood equations

$$\frac{\partial \log L}{\partial a_g} = 0, \qquad \frac{\partial \log L}{\partial b_g} = 0,$$

are easily simplified to

$$\frac{1}{N} \sum_{c=1}^{N} \theta_c \Psi[Da_g(\theta_c - b_g)] = t_g, \qquad g = 1, \ldots, n, \tag{17.9.3}$$

$$\frac{1}{N} \sum_{c=1}^{N} \Psi[Da_g(\theta_c - b_g)] = s_g, \qquad g = 1, \ldots, n, \tag{17.9.4}$$

where

$$s_g = \frac{1}{N} \sum_{c=1}^{N} u_{gc} \qquad \text{and} \qquad t_g = \frac{1}{N} \sum_{c=1}^{N} \theta_c u_{gc}.$$

For each g, the pair of equations (17.9.3) and (17.9.4) in a_g and b_g can be solved for the maximum likelihood estimates $\hat{a}_g$ and $\hat{b}_g$ by numerical iteration with the aid of Berkson's (1957) tables of Ψ.

After each cycle, or after several cycles, of calculation of the successive approximation values

$$[a_g^{(1)}, b_g^{(1)}], \ldots, [a_g^{(r)}, b_g^{(r)}], \qquad g = 1, \ldots, n,$$

the first trial values

$$\theta_c^{(1)} = \theta(x_c) \tag{17.9.5}$$

given by (17.9.2) may be replaced by the successive approximations $\hat{\theta}_c^{(r)}$, for $c = 1, \ldots, n$, where $\hat{\theta}_c^{(r)}$ is a formal solution of the equation for estimation of θ_c when all item parameters are assumed known. This formal solution and its conditions are discussed in detail in Section 20.3 below and used in the next paragraph.

2. Dropping now the assumption made in (1) of a known prior distribution of abilities, we may obtain from L the maximum likelihood estimates $\hat{\theta}_c$ of the examinees' abilities θ_c, along with the estimates $\hat{a}_g$ and $\hat{b}_g$ of item parameters. Even in this case it is convenient to begin an iterative procedure for computing all $\hat{\theta}_c$, $\hat{a}_g$, and $\hat{b}_g$ with first-cycle values $\theta_c^{(1)} = \theta(x_c)$ defined as in (17.9.2). Then second-cycle values $\theta_c^{(2)}$ can be obtained from the maximum likelihood equation (see Section 20.3)

$$\partial \log L/\partial\theta_c = 0,$$

or

$$\sum_{g=1}^{n} a_g\Psi[Da_g(\theta_c - b_g)] = \sum_{g=1}^{n} a_g u_{gc}, \tag{17.9.6}$$

with a_g and b_g replaced by $a_g^{(1)}$ and $b_g^{(1)}$. Then $\hat{\theta}^{(1)}$ can be replaced by $\hat{\theta}^{(2)}$ in (17.9.3) and (17.9.4), and the second-cycle values $a_g^{(2)}$ and $b_g^{(2)}$ can be obtained as solutions of those equations. Further cycles could run through (17.9.6), (17.9.3), and (17.9.4) in several possible patterns of iteration.

Lord (1967) has successfully applied a procedure similar to that just outlined to various sets of data, using a computer program written by Diana Lees. In one application, the a_g, b_g, and θ_c values were simultaneously estimated for 3000 examinees and 90 items (a total of 270,000 item responses). Bock (1967) has reported successful estimation of a_g and b_g values by a method based on the assumption that θ_c is normally distributed in the population of examinees. Substantial variation in a_g values was found in both of these applications.

17.10 Validity of Test Models

Some aspects of questions of validity and adequacy of fit of specific test models were discussed in Chapter 16. For the logistic model, the estimation methods indicated above may be useful as part of an empirical test of fit. Where specific techniques of testing fit are concerned, the reader should be aware that some established approaches to testing goodness of fit have come to be considered unsound and potentially misleading by a number of statisticians and scientific workers. An alternative perspective on testing adequacy of models is one based primarily on rather direct, often graphical, comparisons of data with significant aspects of models. Here a crucial role is played by relatively unformalized judgments that involve both the subject-matter context and statistical considerations. Bush (1963) has described and illustrated one such perspective on testing models.

The bearing of some of these questions on statistical efficiency of estimation of ability will be discussed in Section 19.1.

References and Selected Readings

BERKSON, J., A statistically precise and relatively simple method of estimating the bio-assay with quantal response, based on the logistic function. *Journal of the American Statistical Association,* 1953, **48,** 565–599.

BERKSON, J., Tables for the maximum likelihood estimate of the logistic function. *Biometrics,* 1957, **13,** 28–34.

BIRNBAUM, A., Efficient design and use of tests of a mental ability for various decision-making problems. *Series Report No. 58–16.* Project No. 7755–23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas, January 1957.

BIRNBAUM, A., On the estimation of mental ability. *Series Report No. 15.* Project No. 7755–23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas, 1958. (a)

BIRNBAUM, A., Further considerations of efficiency in tests of a mental ability. *Technical Report No. 17.* Project No. 7755–23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas, 1958. (b)

BIRNBAUM, A., Statistical theory of some quantal response models. *Annals of Mathematical Statistics,* 1958, **29,** 1284 (abstract). (c)

BIRNBAUM, A., Statistical theory of tests of a mental ability. *Op. cit.,* 1285 (abstract). (d)

BIRNBAUM, A., A unified theory of estimation, I. *Annals of Mathematical Statistics,* 1961, **32,** 112–135. (a)

BIRNBAUM, A., The theory of statistical inference. New York: Institute of Mathematical Sciences, New York University, 1961. (b) (Mimeographed)

BIRNBAUM, A., Statistical theory for logistic mental test models with a prior distribution of ability. *Research Bulletin 67–12.* Princeton, N.J.: Educational Testing Service, 1967.

BOCK, R. D., Fitting a response model for *n* dichotomous items. Paper read at the Psychometric Society Meeting, Madison, Wisconsin, March 1967.

BUSH, R. B., *Handbook of mathematical psychology,* Vol. 1, Chapter 8: Estimation and evaluation. New York: Wiley, 1963.

Cochran, W. G., Estimation of bacterial densities by means of the most probable number. *Biometrics*, 1950, **6,** 105–116.

Cramér, H., *Mathematical methods of statistics.* Princeton, N.J.: Princeton University Press, 1946.

Finney, D. J., The application of probit analysis to the results of mental tests. *Psychometrika*, 1944, **9,** 31–39.

Finney, D. J., *Probit analysis.* London: Cambridge University Press, 1952.

Fisher, R. A., On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London* (A), 1922, **222,** 309–368. (Reprinted in R. A. Fisher, *Contributions to mathematical statistics.* New York: Wiley, 1950.)

Guttman, L., Chapters 2, 3, 6, 8, 9 in S. A. Stouffer *et al.*, *Measurement and prediction.* Princeton, N.J.: Princeton University Press, 1950.

Haley, D. C., Estimation of the dosage mortality relationship when the dose is subject to error. *Technical Report No. 15*, August 29, 1952. Stanford, Calif.: Contract No. ONR–25140, Applied Mathematics and Statistics Laboratory, Stanford University.

Kempthorne, O., *An introduction to genetic statistics.* New York: Wiley, 1957.

Lazarsfeld, P., Latent structure analysis. In S. Koch (Ed.), *Psychology: a study of a science*, Vol. 3. New York: McGraw-Hill, 1959.

Lindgren, B. W., *Statistical theory.* New York: Macmillan, 1960, 1962.

Loève, M., *Probability theory.* New York: Van Nostrand, 1955.

Lord, F. M., A theory of test scores. *Psychometric Monograph, No. 7.* Chicago: University of Chicago Press, 1952. (a)

Lord, F. M., The relation of the reliability of multiple-choice tests to the distribution of item difficulties. *Psychometrika*, 1952, **17,** 181–194. (b)

Lord, F. M., An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. *Psychometrika*, 1953, **18,** 57–76.

Lord, F. M., An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Research Bulletin 67–34.* Princeton, N.J.: Educational Testing Service, 1967.

McNemar, Q., *Psychological statistics.* New York: Wiley, 1962.

Rao, C. R., *Linear statistical inference and its applications.* New York: Wiley, 1965.

Rasch, G., *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Nielson and Lydiche (for Danmarks Paedagogiske Institut), 1960.

Stevens, W. L., Control by gauging. *Journal of the Royal Statistical Society* (B), 1948, **10,** 54–108.

Torgerson, W. S., *Theory and methods of scaling.* New York: Wiley, 1958.

Tucker, L. R, Maximum validity of a test with equivalent items. *Psychometrika*, 1946, **11,** 1–14.

Tucker, L. R, Academic ability test. *Research Memorandum 51–17.* Princeton, N.J.: Educational Testing Service, 1951.

If $\varphi(x)$ is the distribution of X, then

$$\varphi(x) = \int_{-\infty}^{\infty} g(x - e)p(e)\, de. \tag{22.6.5}$$

The conditional distribution of E for given x is now

$$f(e \mid x) = \frac{g(x - e)p(e)}{\varphi(x)}\,. \tag{22.6.6}$$

The conditional mean of E for given x is thus

$$\mu(E \mid x) = \frac{1}{\varphi(x)} \int_{-\infty}^{\infty} eg(x - e)p(e)\, de. \tag{22.6.7}$$

Integrating by parts and using (22.6.2), we obtain

$$\begin{aligned}\mu(E \mid x) &= \frac{1}{\varphi(x)}\left[-\sigma^2 g(x - e)p(e) + \sigma^2 \int p(e)g'_e(x - e)\, de \right]_{-\infty}^{+\infty} \\ &= \frac{\sigma^2}{\varphi(x)} \int_{-\infty}^{\infty} p(e)g'_e(x - e)\, de,\end{aligned} \tag{22.6.8}$$

where

$$g'_e(x - e) \equiv \frac{\partial}{\partial e}\, g(x - e). \tag{22.6.9}$$

The first term in brackets in (22.6.8) vanishes because of the behavior of $p(e)$ at $e = \pm\infty$. The integration by parts is permissible if the last integral in (22.6.8) exists; in particular, if g' exists and is bounded.

Now

$$\frac{\partial}{\partial x}\, g(x - e) = -g'_e(x - e). \tag{22.6.10}$$

Substitute this into (22.6.8), and use (22.6.5) to find

$$\mu(E \mid x) = -\frac{\sigma^2}{\varphi(x)} \frac{\partial}{\partial x} \int_{-\infty}^{\infty} p(e)g(x - e)\, de = -\sigma^2 \frac{\varphi'(x)}{\varphi(x)}\,, \tag{22.6.11}$$

where

$$\varphi'(x) \equiv \frac{d}{dx}\, \varphi(x). \tag{22.6.12}$$

This last step, again, is permissible so long as g' exists and is bounded.

The conditional mean of the true score for a given observed score, finally, is

$$\mu(\mathrm{T} \mid x) = \mu(X - E \mid x) = \mu(X \mid x) - \mu(E \mid x) = x + \sigma^2[\varphi'(x)/\varphi(x)]. \tag{22.6.13}$$

Thus the regression of true score on observed score is a simple function of the frequency distribution of the observed score, the slope of this distribution, and the variance of the errors of measurement. If known, this regression can be used to estimate any examinee's true score from his observed score.

Note that *if the distribution of X is unimodal, the estimated true score* $\mu(\mathrm{T} \mid x)$ *for any examinee whose observed score is below the mode of the distribution of X will be higher than his observed score; the estimated true score for any observed score above the mode will be less than the observed score.* This effect might be called *regression toward the mode.* It may be summarized roughly by stating that *extreme observed scores should be somewhat discounted as probably attributable in part to extreme errors of measurement.*

If the distribution of observed scores is flat within any interval of X, there will be no regression toward the mode for observed scores in this interval. Note, however, that if the errors are normally distributed with fixed variance $\sigma^2 > 0$, and if (22.6.13) is valid, then the distribution of X cannot be rectangular. We may prove this by supposing the converse. If the distribution were rectangular, then the regression of true score on observed score would be $\mu(\mathrm{T} \mid x) = x$, by (22.6.13), and would coincide with the regression of observed score on true score (3.7.1a). However, two linear regressions can coincide only if the two variables (X and T) are perfectly linearly related. This cannot occur if $\sigma^2 > 0$.

A formula for $\mathscr{E}\{[\mathrm{T}-\mu(\mathrm{T} \mid x)]^2 \mid x\}$, the variance of the errors of prediction when predicting T from x, could presumably be derived by an extension of the derivation already given. This is left as an exercise for the interested reader.

In practical work (see Trumpler and Weaver), it is necessary to substitute the first differences of the sample grouped frequency distribution for the derivative required in (22.6.13). To the authors' best knowledge, no one has successfully studied the statistical inference problems involved and found a practical modification of (22.6.13) that will provide a consistent and, if possible, reasonably efficient and unbiased estimator of $\mu(\mathrm{T} \mid x)$.

The lack of such an estimator is not too serious if, as is probably often true, the regression $\mu(\mathrm{T} \mid x)$ is not too nonlinear. In such cases, the linear least-squares regression of Section 3.7 can be used.

22.7 The Assumption of Normally Distributed Errors

Pollard (1953) gives a mathematical condition on $\varphi(x)$, the frequency distribution of X, that is both sufficient and necessary for the errors to be normally distributed with constant variance for fixed T. Since Pollard's condition involves an infinite number of derivatives of $\varphi(x)$, it does not readily lend itself to statistical inference. A similar comment applies to a multivariate extension of Pollard's result given by Standish (1956). No adequate alternative practical procedures appear to be available.

Strictly speaking, if the test score X is the number of items answered correctly, it is clear that E cannot be normally distributed for fixed T, since in this case X is both discrete and bounded. If the test score is neither discrete nor bounded, the assumption of normally distributed errors may be plausible.

At the end of Chapter 10, we summarized a study by Lord (1960) which showed that if the test score is taken as the number of items answered correctly, then the errors of measurement studied were not normally distributed, nor were they distributed independently of true score. In Chapter 23, we shall discuss a possible model for the distribution of the errors of measurement in such cases.

22.8 Conditions for Linear Regression of True Score on Observed Score

If (22.6.13) gives the regression of true score on observed score, then the condition for linearity (assuming that $\sigma^2 \neq 0$) is that the last term be a linear function of x; that is, the condition is

$$\frac{\varphi'(x)}{\varphi(x)} = A + Bx, \tag{22.8.1}$$

where A and B are unknown constants. Integrate both sides to obtain

$$\log \varphi(x) = Ax + \tfrac{1}{2}Bx^2 + C,$$

or

$$\varphi(x) = \exp\left(\tfrac{1}{2}Bx^2 + Ax + C\right). \tag{22.8.2}$$

Since $\varphi(x)$ is a frequency distribution, (22.8.2) shows that the observed score x must be normally distributed. Since the errors are normally distributed, independently of true score, it follows that the true score must be normally distributed also. Thus we have

> **Theorem 22.8.1.** *If the errors of measurement are normally distributed independently of true score* (as assumed in 22.6), *then, under the regularity conditions assumed in Section 22.6, the regression of true score on observed score will be linear if and only if the true scores are normally distributed.*

If a particular population of examinees has a normal distribution of true scores, but a subpopulation is selected in which this distribution is not normal, then the regression of true score on observed score in the subpopulation will be nonlinear.

A general result can be obtained without assuming any particular form for the frequency distribution of the errors of measurement.* Kendall and Stuart (1961, Section 28.5) have shown that if some variable Y_2 has a linear regression on Y_1, so that $\mathscr{E}(Y_2 \mid Y_1) = A + BY_1$, then

$$\left.\frac{\partial\psi(\theta_1, \theta_2)}{\partial\theta_2}\right|_{\theta_2=0} = iA + B\,\frac{d\psi(\theta_1, 0)}{d\theta_1}, \tag{22.8.3}$$

where $i \equiv \sqrt{-1}$ and $\psi(\theta_1, \theta_2)$ is the bivariate cumulant generating function of Y_1 and Y_2.

* The remainder of this section assumes some familiarity with characteristic functions. The reader may skip to the next section without loss of continuity.

This result can be applied to the present problem by starting with the bivariate characteristic function of X and T, which may be written

$$\begin{aligned}\Phi(\theta_1, \theta_2) &\equiv \mathscr{E} \exp\,(i\theta_1 X + i\theta_2 \mathrm{T}) = \mathscr{E} \exp\,[i\theta_1(\mathrm{T} + E) + i\theta_2 \mathrm{T}] \\ &= \mathscr{E}\{\exp\,[i(\theta_1 + \theta_2)\mathrm{T}] \exp\,(i\theta_1 E)\} = \Phi_{\mathrm{T}}(\theta_1 + \theta_2)\Phi_E(\theta_1), \qquad (22.8.4)\end{aligned}$$

where $\Phi_{\mathrm{T}}(\theta)$ and $\Phi_E(\theta)$ are the characteristic functions of true score and error, respectively. Thus, the cumulant generating function for observed score and true score can be written

$$\psi(\theta_1, \theta_2) = \psi_{\mathrm{T}}(\theta_1 + \theta_2) + \psi_E(\theta_1). \qquad (22.8.5)$$

Differentiate (22.8.5) to obtain

$$\begin{aligned}\frac{\partial\psi(\theta_1, \theta_2)}{\partial\theta_1} &= \left.\frac{d\psi_{\mathrm{T}}(\theta)}{d\theta}\right|_{\theta=\theta_1+\theta_2} + \left.\frac{d\psi_E(\theta)}{d\theta}\right|_{\theta=\theta_1}, \\ \frac{\partial\psi(\theta_1, \theta_2)}{\partial\theta_2} &= \left.\frac{d\psi_{\mathrm{T}}(\theta)}{d\theta}\right|_{\theta=\theta_1+\theta_2}.\end{aligned}$$

Insert these results into (22.8.3) to obtain a necessary condition for linearity of regression:

$$\left.\frac{d\psi_{\mathrm{T}}(\theta)}{d\theta}\right|_{\theta=\theta_1} = iA + B\left[\frac{d\psi_{\mathrm{T}}(\theta)}{d\theta} + \frac{d\psi_E(\theta)}{d\theta}\right]_{\theta=\theta_1}.$$

Without loss of generality, we can suppose that a constant has been subtracted from all observed scores so that the mean of X in the population of examinees is zero. Thus $A = 0$ when $Y_2 =$ true score and $Y_1 =$ observed score. Also, by (3.7.2), $B = \rho$, where ρ is the test reliability coefficient, which we assume to be nonzero. The last displayed equation can now be written

$$(1 - \rho)\frac{d\psi_{\mathrm{T}}(\theta)}{d\theta} = \rho\frac{d\psi_E(\theta)}{d\theta}. \qquad (22.8.6)$$

Integrate both sides to obtain, finally,

$$(1 - \rho)\psi_{\mathrm{T}}(\theta) = \rho\psi_E(\theta). \qquad (22.8.7)$$

(Any constants of integration must cancel out, since $\psi(0) = 0$ for any random variable.) Thus, *when the errors are distributed independently of true score, a necessary condition for true score to have a linear regression on observed score is that $\psi_{\mathrm{T}}(\theta)$, the cumulant generating function of the true scores, be a constant multiple of $\psi_E(\theta)$, the cumulant generating function of the errors of measurement.*

If the cumulant generating function can be expanded in a power series, then

$$\left.\frac{d^r\psi(\theta)}{d\theta^r}\right|_{\theta=0}$$

is the rth cumulant. Thus, under regularity conditions, (22.8.7) shows that

$$(1 - \rho)\kappa_{\mathrm{T}}^{(r)} = \rho\kappa_E^{(r)}, \qquad r = 2, 3, \ldots . \tag{22.8.8}$$

Taken together with (22.2.4), this shows that the true-score cumulants after the first are the same as the observed-score cumulants, except for the constant factor ρ:

$$\kappa_{\mathrm{T}}^{(r)} = \rho\kappa_X^{(r)}, \qquad r = 2, 3, \ldots . \tag{22.8.9}$$

A special case satisfying (22.8.7), (22.8.8), and (22.8.9) is, of course, the case where the true scores and the errors of measurement are both normally distributed. Here

$$\psi_{\mathrm{T}}(\theta) = -\tfrac{1}{2}\sigma_{\mathrm{T}}^2\theta^2, \quad \psi_E(\theta) = -\tfrac{1}{2}\sigma_E^2\theta^2. \tag{22.8.10}$$

The reader may verify (22.8.8) and (22.8.7) for this special case, remembering that $\rho = \sigma_{\mathrm{T}}^2/\sigma_X^2$ and $\sigma_X^2 = \sigma_{\mathrm{T}}^2 + \sigma_E^2$.

22.9 Conditions for Linear Multiple Regression of True Score on Two or More τ-Equivalent Observed Scores

Ferguson (1955, Theorem 5) has proved a theorem that can be specialized for present purposes.

Theorem 22.9.1. *Let the random variables $X_1, X_2, \ldots, X_g, \ldots, X_n$, $n \geq 2$, be τ-equivalent measurements on a population of examinees, the errors of measurement $X_g - \mathrm{T}$, $g = 1, 2, \ldots, n$, being distributed independently of each other and of the true score* T *with zero means and nonzero variances. Then for the multiple regression of* T *on $X_1, \ldots, X_n$ to be linear, it is necessary and sufficient that* T, *$X_1, \ldots, X_n$ be jointly normally distributed.*

The proof of the theorem will not be given here.

It appears from this theorem that strict linear regression of true score on observed score is a rather specialized and unusual situation.

22.10 Conditions for Linear Regression of One Measurement on Another

Lindley (1947, Section 3.1) has proved a theorem that includes the results of Section 22.8 as special cases. Rephrased for present purposes, it is

Theorem 22.10.1. *Given that*

a) *true score T_1 has a linear regression on true score T_2 with slope β, and*

b) *the errors of measurement $E_1 = X_1 - \mathrm{T}_1$ and $E_2 = X_2 - \mathrm{T}_2$ are distributed independently of each other and of T_1 and T_2,*

then a necessary and sufficient condition for the regression of observed score X_1 on observed score X_2 to be linear is that the cumulant generating function of T_2 be a multiple of the cumulant generating function of E_2. Specifically, this *condition is*

$$(\beta - B)\psi_{T_2}(\theta) = B\psi_{E_2}(\theta), \tag{22.10.1}$$

where B is the slope of the regression of X_1 on X_2.

A proof of this result is also given by Kendall and Stuart (1961, Section 29.57).

Lindley has given several related theorems, including the multivariate generalization of Theorem 22.10.1. The interested reader is referred to his article and to Ferguson's.

It appears from this theorem that strict linear regression of one observed score on another is a specialized occurrence. However, approximate linear regression does seem to hold in practice for many sets of empirical data.

Exercise

22.1. Suppose X and Y are essentially τ-equivalent. Show that the regression of X on Y is linear if and only if the regression of true score on observed score is linear for Y.

References and Selected Readings

EDDINGTON, A. S., On a formula for correcting statistics for the effects of a known probable error of observation. *Royal Astronomical Society Monthly Notices*, 1913, **73,** 359–360.

FERGUSON, T., On the existence of linear regression in linear structural relations. *University of California Publications in Statistics*, 1955, **2,** No. 7.

GAFFEY, W. R., A consistent estimator of a component of a convolution. *Annals of Mathematical Statistics*, 1959, **30,** 198–205.

GIRSHICK, M. A., and L. J. SAVAGE, Bayes and minimax estimates for quadratic loss functions. In J. Neyman (Ed.), *Proceedings of the second Berkeley symposium on mathematical statistics and probability.* Berkeley: University of California Press, 1951, pp. 53–73.

HIRSCHMAN, I. I., and D. V. WIDDER, *The convolution transform.* Princeton, N.J.: Princeton University Press, 1955.

KENDALL, M. G., and A. STUART, *The advanced theory of statistics.* Vol. 1: *Distribution theory.* New York: Hafner, 1958.

KENDALL, M. G., and A. STUART, *The advanced theory of statistics.* Vol. 2: *Inference and relationship.* New York: Hafner, 1961.

KURTH, R., On Eddington's solution of the convolution integral equation. Rendiconti del Circolo Matematico di Palermo, 1965, **14,** 76–84.

LINDLEY, D. V., Regression lines and the linear functional relationship. *Journal of the Royal Statistical Society*, 1947, **9,** 218–244.

LORD, F. M., An empirical study of the normality and independence of errors of measurement in test scores. *Psychometrika*, 1960, **25,** 91–104.

NOVICK, M. R., The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, 1966, **3,** 1–18.

NOVICK, M. R., and C. LEWIS, Coefficient alpha and the reliability of composite measurements. *Psychometrika*, 1967, **32,** 1–13.

PATIL, G. P. (Ed.), *Classical and contagious distributions.* New York: Pergamon Press, 1965.

PATIL, G. P., and S. W. JOSHI, Bibliography of classical and contagious discrete distributions. ARL 66–0185, Aerospace Research Laboratories, U.S. Air Force, September 1966.

PITMAN, E. J. G., The estimation of the location and scale parameters of a continuous population of any given form. *Biometrika*, 1939, **30,** 391–421.

POLLARD, H., Distribution functions containing a Gaussian factor. *Proceedings of the American Mathematical Society*, 1953, **4,** 578–582.

ROBBINS, H., An empirical Bayes approach to statistics. In J. Neyman (Ed.), *Proceedings of the third Berkeley symposium on mathematical statistics and probability*, Vol. 5. Berkeley: University of California Press, 1956, pp. 157–163.

ROBBINS, H., The empirical Bayes approach to testing statistical hypotheses. *Review of the International Statistical Institute*, 1963, **31,** 195–208.

ROBBINS, H., The empirical Bayes approach to statistical decision problems. *Annals of Mathematical Statistics*, 1964, **35,** 1–20.

SCARBOROUGH, J. B., *Numerical mathematical analysis*, 3rd ed. Baltimore: The Johns Hopkins Press, 1955.

STANDISH, C., *N*-dimensional distributions containing a normal component. *Annals of Mathematical Statistics*, 1956, **27,** 1161–1165.

TEICHER, H., On the mixture of distributions. *Annals of Mathematical Statistics*, 1960, **31,** 55–73.

TRUMPLER, R. J., and H. F. WEAVER, *Statistical astronomy.* Berkeley: University of California Press, 1953.

9 781593 119348