When populations are cross-classified with respect to two or more classifications or polytomies, questions often arise about the degree of association existing between the several polytomies. Most of the traditional measures or indices of association are based upon the standard chi-square statistic or on an assumption of underlying joint normality. In this paper a number of alternative measures are considered, almost all based upon a probabilistic model for activity to which the cross-classification may typically lead. Only the case in which the population is completely known is considered, so no question of sampling or measurement error appears. We hope, however, to publish before long some approximate distributions for sample estimators of the measures we propose, and approximate tests of hypotheses. Our major theme is that the measures of association used by an empirical investigator should not be blindly chosen because of tradition and convention only, although these factors may properly be given some weight, but should be constructed in a manner having operational meaning within the context of the particular problem.

## 1. INTRODUCTION

Many studies, particularly in the social sciences, deal with populations of individuals which are thought of as cross-classified by two or more polytomies. For example, the adult individuals living in New York City may be classified as to

| | |
|---|---|
| Borough: | 5 classes |
| Newspaper most often read: | perhaps 6 classes |
| Television set in home or not: | 2 classes |
| Level of formal education: | perhaps 5 classes |
| Age: | perhaps 10 classes |

For simplicity we deal largely with the case of two polytomies, although many of our remarks may be extended to a greater number. The double polytomy is the most common, no doubt because of the ease with which it can be tabulated and displayed on the printed page. Most of our remarks suppose the population completely known in regard to the classifications, and indeed this seems to be the way to begin in the construction of rational measures of association. After agreement has been reached on the utility of a measure for a known population, then

(London School of Economics and Political Science), Frederick Mosteller (Harvard University), I. Richard Savage (National Bureau of Standards), Alan Stuart (London School of Economics and Political Science), Louis L. Thurstone (University of North Carolina), John W. Tukey (Princeton (University), W. Allen Wallis (University of Chicago), and E. J. Williams (Commonwealth Scientific and Industrial Research Organization, Australia). Part of Mr. Goodman's work on this paper was carried out at the Statistical Laboratory of the University of Cambridge under a Fulbright Award and a Social Science Research Council Fellowship. The authors were led to work on the problems of this paper as a result of conversations with Louis L. Thurstone and Bernard R. Berelson.

one should consider the sampling problems associated with estimation and tests about this population parameter.

A double polytomy may be represented by a table of the following kind:[1]

| $A$ | $B$ | | | | |
|---|---|---|---|---|---|
| | $B_1$ | $B_2$ | $\cdots$ | $B_\beta$ | Total |
| $A_1$ | $p_{11}$ | $p_{12}$ | $\cdots$ | $p_{1\beta}$ | $p_{1\cdot}$ |
| $A_2$ | $p_{21}$ | $p_{22}$ | $\cdots$ | $p_{2\beta}$ | $p_{2\cdot}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $A_\alpha$ | $p_{\alpha 1}$ | $p_{\alpha 2}$ | $\cdots$ | $p_{\alpha\beta}$ | $p_{\alpha\cdot}$ |
| Total | $p_{\cdot 1}$ | $p_{\cdot 2}$ | $\cdots$ | $p_{\cdot\beta}$ | 1 |

where

Classification $A$ divides the population into the $\alpha$ classes $A_1, A_2, \cdots, A_\alpha$.

Classification $B$ divides the population into the $\beta$ classes $B_1, B_2, \cdots, B_\beta$.

The proportion of the population that is classified as both $A_a$ and $B_b$ is $p_{ab}$.

The marginal proportions will be denoted by

$p_{a\cdot}=$ the proportion of the population classified as $A_a$.
$p_{\cdot b}=$ the proportion of the population classified as $B_b$.

If the use to which a measure of association were to be put could be precisely stated, there would be little difficulty in defining an appropriate measure. For example, using the above cross-classification of the New York City population, a television service company might wish to

[1] Tables of this kind are frequently called *contingency tables*. We shall not use this term because of its connotation of a specific sampling scheme when the population is not known and one infers on the basis of a sample.

There are vaguenesses in the idea of complete ordered association. For example, everyone would probably agree that the following case is one of complete association:

| | | |
|---|---|---|
| 0 | 0 | 0 |
| $\rho_{21}$ | 0 | 0 |
| 0 | $\rho_{32}$ | 0 |

The following situation is not so clear:

| | | |
|---|---|---|
| $\rho_{11}$ | 0 | 0 |
| $\rho_{21}$ | $\rho_{22}$ | 0 |
| 0 | $\rho_{32}$ | $\rho_{33}$ |

As before, the procedure we shall adopt toward this and toward more complex questions is to base the measure of association on a probabilistic model of activity which often may be appropriate and typical.

### 6.2. *A Proposed Measure*

Our proposed model will now be described. Suppose that two individuals are taken independently and at random from the population (technically with replacement, but this is unimportant for large populations). Each falls into some $(A_a, B_b)$ cell. Let us say that the first falls in the $(A_{\underline{a}_1}, B_{\underline{b}_1})$ cell, and the second in the $(A_{\underline{a}_2}, B_{\underline{b}_2})$ cell. (Underlined letters denote random variables.) $\underline{a}_i$ $(i=1, 2)$ takes values from 1 to $\alpha$; $\underline{b}_i$ $(i=1, 2)$ takes values from 1 to $\beta$.

If there is independence, one expects that the order of the $\underline{a}$'s has no connection with the order of the $\underline{b}$'s. If there is high association one expects that the order of the $\underline{a}$'s would generally be the same as that of the $\underline{b}$'s. If there is high counterassociation one expects that the orders would generally be different.

Let us therefore ask about the probabilities for like and unlike orders. In order to avoid ambiguity, these probabilities will be taken conditionally on the absence of ties. Set

$$\text{(18)}\quad \Pi_s = \Pr\{\underline{a}_1 < \underline{a}_2 \text{ and } \underline{b}_1 < \underline{b}_2; \text{ or } \underline{a}_1 > \underline{a}_2 \text{ and } \underline{b}_1 > \underline{b}_2\}$$

$$\text{(19)}\quad \Pi_d = \Pr\{\underline{a}_1 < \underline{a}_2 \text{ and } \underline{b}_1 > \underline{b}_2; \text{ or } \underline{a}_1 > \underline{a}_2 \text{ and } \underline{b}_1 < \underline{b}_2\}$$

$$\text{(20)}\quad \Pi_t = \Pr\{\underline{a}_1 = \underline{a}_2 \text{ or } \underline{b}_1 = \underline{b}_2\}.$$

Then the conditional probability of like orders given no ties is $\Pi_s/(1-\Pi_t)$ and the conditional probability of unlike orders given no ties is $\Pi_d/(1-\Pi_t)$. Of course, the sum of these two quantities is one.

A possible measure of association would then be $\Pi_s/(1-\Pi_t)$, but it is a bit more convenient to look at the following quantity:

$$\text{(21)}\qquad \gamma = \frac{\Pi_s - \Pi_d}{1 - \Pi_t}$$

or the *difference* between the conditional probabilities of like and unlike orders. In other words $\gamma$ tells us how much more probable it is to get like than unlike orders in the two classifications, when two individuals are chosen at random from the population.

Since $\Pi_s + \Pi_d = 1 - \Pi_t$, we may write $\gamma$ as

$$\text{(22)}\qquad \gamma = \frac{2\Pi_s - 1 + \Pi_t}{1 - \Pi_t}$$

which is convenient for computation, using the easily checked relationships

$$\text{(23)}\qquad \Pi_s = 2\sum_a \sum_b \rho_{ab}\Big\{\sum_{a'>a} \sum_{b'>b} \rho_{a'b'}\Big\}$$

$$\text{(24)}\qquad \Pi_t = \sum_a \rho_{a\cdot}^2 + \sum_b \rho_{\cdot b}^2 - \sum_a \sum_b \rho_{ab}^2.$$

Some important properties of $\gamma$ follow:

(*i*) $\gamma$ is indeterminate if the population is concentrated in a single row or column of the cross-classification table.

(*ii*) $\gamma$ is 1 if the population is concentrated in an upper-left to lower-right diagonal of the cross-classification table. $\gamma$ is $-1$ if the population is concentrated in a lower-left to upper-right diagonal of the table.

(*iii*) $\gamma$ is 0 in the case of independence, but the converse need not hold except in the $2\times 2$ case. An example of nonindependence with $\gamma = 0$ is