# Significance Testing in a Nonstochastic Setting

by

David Freedman, University of California, Berkeley

and

David Lane, University of Minnesota

Technical Report No. 317

## 1. Introduction

> "The function of tests of significance is to prevent us being deceived by accidental occurrences, due not to causes we wish to study, or are trying to detect, but to a combination of the many other circumstances which we cannot control."
>
> R. A. Fisher (1929, p. 21).

The conventional theory of significance tests can be used only when the assumptions which underlie it apply. As we shall indicate by examples, tests are routinely carried out in situations where these assumptions are not satisfied. Significance probabilities are then very difficult to interpret. Our object in this paper is to sketch an alternative interpretation of significance probabilities which often makes sense even when the assumptions of the conventional theory do not hold.

The conventional theory of significance testing applies if, for example, data is collected in a random sample, or subjects are assigned to treatments by randomization, or the investigators have a reasonable random model for the mechanism producing their data. The theory requires that the data be embedded in a stochastic framework, complete with random variables, probability distributions, and unknown parameters. However, as our examples in section 2 illustrate, the data often arrive without benefit of randomness. In such cases, the investigators may still wish to separate effects of "the causes they wish to study or are trying to detect" from "accidental occurrences due to the many other circumstances which they cannot control." What can they do? Usually, they follow Fisher (1922) into a fantasy world "by constructing a hypothetical infinite population, of which the actual data are regarded as constituting a random sample." Unfortunately, this fantasy world is often harder to understand than the original problem which lead to its invocation.

Section 2 begins with two typical examples taken from the statistical literature. In these examples, a $\chi^2$-test and an F-test are carried out. On first reading these examples, many statisticians will find them perfectly routine and straightforward. However, we shall review the conventional theory of significance testing and show that it does not apply to these examples. The section concludes with our attempt to make sense out of the significance probabilities calculated in the tests of our two examples.

While discussing the ideas of this paper with colleagues, we have frequently encountered two responses. Some colleagues, especially those who work as consulting statisticians, have told us that our approach describes the way they have always thought about testing. Nonetheless, we were unable to find this approach developed in any account of the theory of testing, and so we felt it was worth writing down. Other colleagues told us that "it's all in Fisher," and while that is generally a safe remark about nearly any statistical idea, we were unable to find exactly our point of view in Fisher's works. What we did find there relevant to our theme is described in section 3. Section 3 also contains some notes on related ideas of other authors. In section 4, we present some calculations which support the interpretation of section 2.

## 2. A nonstochastic interpretation of significance probabilities.

Our discussion will focus on two simple, but typical, examples.

Example 1: The data displayed below were collected in the course of an investigation into the existence of sex bias in admissions to

graduate school (see Bickel, Hammel, and O'Connel (1975) for the setting; the data comes from the Graduate Division of the University of California). The figures give the number in each category among all applicants for admission to one of the University of California at Berkeley's six largest departments for the 1973-4 academic year:

| | Accepted | Rejected | Total |
|---|---|---|---|
| Men | 54 | 137 | 191 |
| Women | 94 | 299 | 393 |
| Total | 148 | 436 | 584 |

The acceptance rate for men applicants was 28%; for women, 24%. At this stage of the investigation, before any detailed analysis of the admissions procedure was undertaken, the question arose: can this difference in rates be explained as accidental, or is it real?

A standard answer to this question involves carrying out a $\chi^2$-test for independence. In this case, the significance probability is about 0.26, suggesting that the difference could well be accidental.

Example 2: "Scientific mass appraisal" involves the application of linear regression techniques to the problem of property valuation (Renshaw (1958)). The sale price of a house is to be predicted from known characteristics (taxes, number of baths, lot size, etc.). Narula and Wellington (1977) present and analyze data on 28 homes taken from Multiple Listing, Vol. 87, for Erie, Pa. This data has also been analyzed by Weisberg (1977). It is presented in the appendix.

A standard question which arises in "scientific mass appraisal" is about which of the available characteristics should be included in the prediction equation. Concretely (focussing on two models discussed

by Weisberg): if one knows the amount of taxes paid on a house, and how much living space the house affords, need the number of baths enter the equation for predicting selling price? Assume that the investigators have, by examining their data, convinced themselves that the linear models involved provide a reasonable fit to the data. Then a standard procedure used to answer this question is the F-test. Here, the appropriate F-test yields a significance probability of about 0.10, suggesting that the number of baths does not belong in the equation.

These examples share two features:

(1) The investigators played a passive role in the data collection procedure. Random sampling was not used to determine which cases entered the study, nor was randomization employed to assign different cases to different treatments during the course of the study.

In the first example, every applicant to the department was included in the study. In such a case, it is often suggested that the actual applicants form a sample from the population of all possible applicants. However, this is a very unsatisfactory solution, since neither the population nor the sampling procedure can be sharply defined. As a result, this suggestion provides no way to calculate meaningful probabilities. Randomization has as little to do with this example as sampling: the "assignment" of sex to applicants was not in the hands of the investigators. Finally, by the time the applicants entered the study, the men and women differed in many important ways besides sex, and these may well be relevant to the decision about admissions to graduate school.

In the second example also, the house to house differences of interest to the investigators were intrinsic properties of the houses, not under the investigators' control except by their choice of which

houses were included in the study. Here, that choice was haphazard: the investigators merely lifted a block of 28 entries from the Multiple Listing. But even had every house in the Multiple Listing been included, the purpose of the study was to apply the resulting equation to predict selling prices for houses yet unsold, for which the value of selling price would consequently be unavailable at the time the study was carried out. Again, it is very difficult to view the houses in the Multiple Listing as a simple random sample--or indeed, any kind of probability sample--from any population of houses.

(2) The purpose of the test was not to evaluate a proposed stochastic model for the physical mechanism which produced the data. Rather, the investigators wanted to determine whether some attribute of the data was too pronounced to be easily explained away as some kind of fluke.

In example 1, the attribute of the data which interests the investigators is the difference in the proportions of male and female applicants accepted to the graduate program. Can the observed difference of 4% be regarded as some sort of artifact, or does it require a more substantial explanation? The object of the $\chi^2$-test is to attempt an answer to the question by converting the difference, via the $\chi^2$-statistic and the $\chi^2_1$-distribution, into a significance probability.

The test itself neither involves nor suggests any reasonable stochastic model for the complicated process of making admissions decisions. It does not even address the question of whether or not a "sex parameter" should be part of such a model, since this could not be determined without simultaneously considering the many other factors affecting the decision process, like the qualifications of the applicants or departmental policies on admissions.

In the second example, the sole aim of "scientific mass appraisal" is to develop a formula for predicting selling price. The technology of

regression analysis is used to fit a prediction equation, but in no sense does it provide a stochastic model for the selling process. The first step in constructing such a model would involve identifying the most important factors affecting selling price. "Scientific mass appraisal", on the other hand, begins with a list of characteristics selected because they can be described numerically and are readily available; such determinants of selling price as location of property and methods of selling and financing are related to these characteristics indirectly or not at all. Linear methods are employed in developing the prediction equation because they are easy to apply, not because some insight into the selling process suggests that "mean selling price" conditional on the values of the characteristics considered should be a linear function of the values. (Of course, it would be foolish to use these methods if residual plots suggested highly nonlinear fits). And so long as prediction errors are generally small, it is neither known nor relevant (to the appraisers) whether they are independent of one another or what their distributions are.

The purpose of the analysis is to find a "best predicting" subset of the characteristics considered. The F-test of example 2, then, is addressed to the question: when the equation for selling price based on taxes and living space alone is computed and applied to the 28 houses in the study, is the resulting sum-of-squares-prediction error sufficiently close to the error obtained by fitting an equation based on taxes, living space, and number of baths to justify appraising selling price on the basis of taxes and living space alone?

Testing situations with these two features--that is, in which randomness is not introduced through sampling cases or assigning cases to treatments, and in which the construction of a reasonable random model for some physical process is not at issue--we shall call non-

stochastic, and we shall suggest a way of interpreting tests in this setting free from stochastic considerations.

First, we briefly review the standard formulation of significance tests, following the survey of Cox (1977). As Cox explains, a significance test is "a procedure for measuring the consistency of data with a null hypothesis." This null hypothesis always has a particular form: it is an assertion about the chance mechanism which generated the data. That is, "we have an observed vector, y, ... and a null hypothesis $H_0$, according to which y is the observed value of a random variable Y," whose distribution belongs to some family, also called $H_0$. The formulation of the significance test is then completed by specifying a test statistic t, "such that the larger is t(y), the stronger is the inconsistency with $H_0$, in the respect under study," and such that the observed level of significance, $p_{obs} = P(t \geq t_{obs}; H_0)$ is (at least approximately) computable (Cox (1977), p. 50). This observed significance level, the chance that an experiment governed by $H_0$ generates a value of t more extreme than $t_{obs}$, serves as "a summary statement of the relation between data [y] and a probability distribution [$H_0$] that might have generated that data" [p. 56]. Inference from the test depends upon the value of $p_{obs}$. If $p_{obs}$ is very small, one is entitled to conclude that the data is not consistent with accepting $H_0$ as a chance model for the mechanism which generated the data.

It is hard to see how to incorporate testing in a nonstochastic setting into this framework. Consider example 1. Clearly the observed level of significance cannot be interpreted from a strict frequentist point of view, since the data derive from the inherently nonrepeatable graduate admissions process of 1973-4. Fisher (1956, third edition) considered this difficulty in his discussion of the problem of determining whether the stars were "distributed at random." He argued that for the

purposes of significance testing, the repetitions--and hence the interpretation of observed level of significance as a probability--need only be "hypothetical" [p. 47]. For his problem it does seem possible to perform the "thought experiment" of recreating the world by a process which locates stars "at random," since $H_0$ in this case does provide a chance model (albeit simplified) for the process which produced the data. But in the case of example 1, if $H_0$ does not specify a model which incorporates known aspects of the complicated admissions process (and, as discussed above, the usual null hypotheses for the $\chi^2$-test for independence do not), the relevant "thought experiment" is elusive. Must we imagine many different ensembles of 584 applicants, each ensemble similar in sex distribution (and perhaps with respect to their qualifications, intellectual interests, and so forth) submitting applications to a department which accepts or rejects them by some process which does not take sex--or any of its covariates--into account? Or should we think of the same group of applicants, with sex identifications somehow reassigned at random, but everything else remaining the same, reconsidered by a "forgetful" department which makes new decisions without regard for the sex of the (newly resexed) applicants? Both suggestions are fanciful and unsatisfactory. Worse, both can <u>at best</u> lead to an inference of the form, "these data are inconsistent with a chance model for their generation--which from the first was known to be inadequate to describe any aspect of the physical process which generated the data!" This situation seems inevitable in the nonstochastic setting. To think of the data as an observation of some random vector $Y$ with distribution $H_0$ requires the data to be connected up with randomness either directly by imposing randomness during data collection (through sampling or randomization), or indirectly, through a convincing analogy between "$Y$ under $H_0$" and the actual physical process which generated the data. In the nonstochastic

setting, neither of these links is available.

These difficulties can be avoided if the standard formulation is set aside, and significance probabilities computed in the nonstochastic setting are interpreted nonstochastically. To see how to go about this, we return to our two examples.

In example 1, the investigators computed a significance probability of 0.26. How can this number be interpreted? Suppose the investigators were to divide the applicants into two groups according to whether they are right- or left-handed, or whether their last names begin with A-L or M-Z, or perhaps just according to some whim (so that no name, like "sex" of "handedness", can even be associated with the division). For each such division of the applicants, into two groups, the investigators could determine what proportion of each group was accepted for admission and calculate the value of the corresponding $\chi^2$-statistic. According to theorem 3 below, the proportion of the divisions for which the $\chi^2$-statistic is greater than $\chi^2_{obs}$ (= 1.29) is approximately $p_{obs}$. Thus, about 25% of these divisions yield a larger value for the $\chi^2$-statistic than the sex division. Clearly, most of these divisions must be regarded as irrelevant to the admissions process. Thus the 4% difference in proportion of male and female applicants accepted for admission does not seem particularly unusual, and it seems plausible to describe the difference as an accidental occurrence, due not to sex bias but to the "many circumstances which we cannot control." On the other hand, had a very small value for $p_{obs}$ been obtained, the sex division would appear unlike most other divisions of the applicants, and it would be hard to explain away the difference in proportion of males and females accepted as a mere fluke.

The investigators may wish to control for such factors as the qualifications and intellectual interests of applicants when they test to see whether the difference in proportion of males and females

accepted is accidental or real (see, for example, Kruskal (1977)). This requires construction of higher-order contingency tables. The significance probability from the resulting $\chi^2$-tests can again be given a nonstochastic interpretation: instead of comparing the division by sex to all possible divisions of the applicants into two groups, comparisons are restricted only to divisions which match the sex division with respect to whatever features of the composition of the groups are of interest to the investigators. (If comparisons are restricted to divisions which have the same numbers in each group as the sex division, the proportion of such divisions with $\chi^2$-value larger than $\chi^2_{obs}$ is precisely the significance probability for Fisher's exact test. If there had been very few males or females among the applicants, the $\chi^2_1$-approximation would not be satisfactory, and the significance probability should be computed using Fisher's exact test.)

The significance probability calculated in example 2 may be interpreted in a similar spirit. The investigators wish to evaluate the informal "null hypothesis" that, given taxes and living space, number of baths is unrelated to selling price. They first obtain the least squares prediction equation for selling price based upon taxes and living space; call the predicted selling price for the $i^{th}$ house $P_i$ and the residual (the difference between $P_i$ and the actual selling price) $R_i$. Let $N_i$ be the number of baths for the $i^{th}$ house. The scatter-plot for N against R shows that the relation between these two variables is approximately linear, and that the spread of the residuals for each "number of bath" level is about the same (see the appendix); thus the relation between N and R is well-described by the correlation coefficient.

The "null hypothesis" thus implies that the correlation between $N$ and $R$ should be about zero. Suppose now that the residuals were rearranged arbitrarily (say by the permutation $\pi$ on $\{1, \ldots, 28\}$) to form a new list $\{R_{\pi(i)}\}_{i \le i \le 28}$. Whatever the relation between $N$ and $R$, there is no reason to expect a relation between $N$ and $R_\pi$ for most permutations $\pi$. If the "null hypothesis" were true, the correlation between $N$ and $R$ should be comparable to the correlation between $N$ and $R_\pi$ for most permutations $\pi$. On the other hand, if the "null hypothesis" is false, the absolute value of the correlation between $N$ and $R$ should be larger than the absolute value of the correlation between $N$ and $R_\pi$ for all but a few unusual permutations $\pi$.

Since the F-statistic used in the test of this example is a re-scaled version of the correlation coefficient between $N$ and $R$, this suggests how to interpret this test from a nonstochastic point of view. To begin with, for each permutation $\pi$ on $\{1, \ldots, 28\}$, a new data set can be found as follows. The taxes paid, living space, and number of baths for each house are left unchanged, but the selling price of house $i$ is changed to a "mock selling price"

$$Y_i^\pi = P_i + R_{\pi(i)} \qquad (1 \le i \le 28) .$$

For this new data set, consider predicting the mock selling price from taxes, living space, and number of baths. An F-statistic can be computed from the new data set to see whether the number of baths should go into the predicting equation. By Theorem 2 below, the proportion of permuted data sets which yield an F-statistic larger than $F_{obs}$ is approximately $p_{obs}$. Thus, for about 10% of the permuted data sets, the F-statistic will be larger than the one obtained from the real

data set. Of course, in the permuted data sets, given taxes and living space, any association between "mock selling price" and number of baths is purely accidental. This suggests that the association in the real data set should be considered accidental, and the number of baths should not go into the prediction equation.

If the plot of N against R had not appeared roughly homoscedestic, it would not make sense to compare the permuted lists $Y^{\pi}$ with the actual selling price. However, if the null hypothesis were true, an arbitrary change of signs applied to each residual would produce lists (of the form $P_i \pm R_i$) which would appear comparable to the actual selling prices. Again, if all possible such lists are considered and the appropriate F-statistic computed for each, $p_{obs}$ is approximately the proportion of the sign changes for which $F > F_{obs}$.

The kind of interpretation provided above for the tests of examples 1 and 2 can be given in general for tests in the nonstochastic setting. Here is an outline for this testing procedure:

1. Data are obtained in a nonstochastic setting, and for some attribute of this data, the question is raised: can this attribute be dismissed as an artifact, or does it require a more substantial explanation? (In example 1, the attribute is the difference in proportion of males and females accepted; in example 2, the correlation between number of baths and the residuals $R_i$).

2. A test statistic is introduced, such that the more pronounced the attribute of (1) is for a data set, the larger the value of the test statistic. (In example 1, the $\chi^2$-statistic; in example 2, the F-statistic).

3. A class of transformations of the data are introduced, which when applied to the original data set result in data sets which possess the attribute of (1) only accidentally, but which are otherwise comparable to the original data. (In example 1, the divisions of the applicants into two groups; in example 2, the rearrangements of the residuals to produce "mock selling prices").

4. The values of the statistic of (2) are calculated for all the transformed data sets of (3), and these values are then ordered to obtain a distribution.

5. The relative position in this distribution of the test-statistic for the original data set is determined.

Note that in this interpretation, "observed level of significance," as calculated in (5), is simply a descriptive statistic. However, this descriptive statistic can be interpreted in the conventional way. Large values of the observed significance level do suggest that as far as the test statistic in (2) is concerned, the original data set is like many of the transformed data sets of (3). These data sets possess the attribute of (1) only by accident, suggesting that the attribute noticed in the original data set can be dismissed as an artifact. On the other hand, small values of the observed significance level make this sort of explanation implausible. If the statistic is in the tail, the data are providing evidence that there is something real to study. If the object of the investigation is to understand some aspect of a real process (as in example 1), the next stage of the study should involve experimentation or model building, and carry the investigation beyond the nonstochastic setting considered here.

## 3. Some Related Ideas

1. Fisher achieved a great breakthrough when he realized the role that randomization could play in separating "accidental occurrences" from real effects. In his classic exposition of the tea-tasting experiment in The Design of Experiments, he showed how presenting the cups to the lady in a random order allowed a test for her claimed testing ability which has a simple null distribution based upon an enumeration of the possible orders for presenting the cups and which allows the investigators to disregard "the many other circumstances" which could not be controlled.

Throughout Design, and in most of his other writings, Fisher insisted that the validity of his enumeration tests rested on "the physical act of randomization." In one instance, however, he argued that a hypothetical randomization could provide a valid basis for inference as well.

> "Let us suppose, for example, that we have measurements of the stature of a hundred Englishmen and a hundred Frenchmen. It may be that the first group are, on the average, an inch taller than the second, although the two sets of heights will overlap widely. If the two groups have been chosen from their respective populations in such a way as not to be random samples of the populations they represent, then an examination of the samples will clearly not enable us to compare these populations; but even if our samples are satisfactory in the manner in which they have been obtained, the further question arises as to whether a difference of the magnitude observed might not have occurred by chance, in samples from populations of the same average height. If the probability of this is considerable, that is, if it would have occurred in fifty, or even ten per cent, of such trials, the difference between our samples is said to be "insignificant." If its probability of occurrence is small, such as one in a thousand, or one in a hundred, or even one in twenty trials, it will usually be termed "significant," and be regarded as providing substantial evidence of an average difference in stature between the two populations sampled...
>
> The simplest way of understanding quite rigorously, yet without mathematics, what the calculations of the test of significance amount to, is to consider what would happen if our two hundred actual measurements were written on cards, shuffled without regard to nationality, and divided at random into two new groups of a hundred each. This division could be done in

> an enormous number of ways, but though the number is enormous it is a finite and a calculable number. We may suppose that for each of these ways the difference between the two average statures is calculated. Sometimes it will be less than an inch, sometimes greater. If it is very seldom greater than an inch, in only one hundredth, for example, of the ways in which the subdivision can possibly be made, the statistician will have been right in saying that the samples differed significantly. For if, in fact, the two populations were homogeneous, there would be nothing to distinguish the particular subdivision in which the Frenchmen are separated from the Englishmen from among the aggregate of the other possible separations which might have been made. Actually, the statistician does not carry out this very simple and very tedious process, but his conclusions have no justification beyond the fact that they agree with those which could have been arrived at by this elementary method."
>
> (Fisher (1936), pp. 58-9).

To Fisher, it is an essential part of this argument that the two samples were in fact drawn randomly from the respective populations. So far as we know, he never applied this kind of interpretation to significance probabilities calculated from data obtained in a purely observational study. Our paper is an extension of his argument to the nonstochastic setting. Probably Fisher did not carry out this extension himself because he was wedded to a principle which prohibited this sort of idea, the principle that probabilities express frequencies in some population.

2. Fisher's "hypothetical randomization" argument has been developed in another direction, leading to the theory of conditional permutation tests (see Cox and Hinkley (1974), pp. 182-202). This theory, unlike our extension of Fisher's idea, stays within the world of "hypothetical infinite populations of which the actual data are regarded as constituting a random sample:" its novelty, in the words of one of its early workers, is that "it frankly starts from the sample and works towards the population instead of the reverse" (Pitman, 1937). It does this by obtaining the null distribution <u>conditionally</u> on some aspect of the configuration of the observed data. These null distributions then turn out to be Fisherian enumeration distributions which do not depend on the exact distribution

undergoing sampling. The theorems giving asymptotic approximations to the null distributions of these tests, like Theorem 1 below, can be applied in the nonstochastic setting as well.

3. Forsythe, Engelmann, Jennrich and May (1973), working in a stochastic context, have proposed that the permutation distribution described above for the F-statistic be used in place of the standard distribution (even though, as Theorem 2 below shows, they are asymptotically equivalent). Since computing such a distribution is prohibitively time-consuming, they suggest sampling a subset of the permutation group to approximate the exact distribution.

4. Following the usage of Fisher and Cox, we have confined the meaning of the words "probability" and "stochastic" to a frequentist framework. From a Bayesian point of view, the enumeration distributions which we use to interpret significance probabilities in the nonstochastic setting may be regarded as probability distributions derived from subjective probability assignments giving equal weight to each of the appropriate transformations of the data. Bruce Hill (1969) has taken a somewhat similar point of view in his analysis of least-squares based on errors "conditionally uniform on spheres."

3. Theorems.

$S_n$ is the set of permutations on $\{1, \ldots, n\}$, and for a list of numbers $\{x_i\}_{1\le i\le n}$, $M_r(\{x_i\}) = \frac{1}{n}\sum_{i=1}^{n} x_i^r$.

Theorem 1. (Wald and Wolfowitz (1944)): For each $n$, $\{x_{ni}\}_{1\le i\le n}$ and $\{y_{ni}\}_{1\le i\le n}$ are lists of numbers satisfying:

(i) $M_1(\{x_{ni}\}_i) = M_1(\{y_{ni}\}_i) = 0$

(ii) $M_2(\{x_{ni}\}_i) = M_2(\{x_{ni}\}_i) = 1$

(iii) for each integer $r$, there is a number $c(r)$ such that (for all $n$)

$$M_r(\{x_{ni}\}_i) \le c(r) ,$$

$$M_r(\{y_{ni}\}_i) \le c(r) .$$

Then the distributions of the lists $\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_{ni}y_{n\pi(i)}\}_{\pi\in S_n}$ converge to the standard normal distribution.

To set up Theorem 2, suppose for each $n$ we have a vector $y^n \in R^n$ and a subspace $Z^n \subset R^n$ generated by $\{z_1^n, \ldots, z_s^n\}$ satisfying:

(i) $M_1\{y^n\} = M_1\{z_j^n\} = 0$ ,

$M_2\{y^n\} = M_2\{z_j^n\} = 1$,

$1 \le j \le s$, $n \ge 1$;

(ii) there exists an $\epsilon > 0$, independent of $n$, such that for each $i$, $1 \le i \le s$, $R^2_{z_1^n, \ldots, z_{i-1}^n \cdot z_i^n} < 1-\epsilon$, and $R^2_{z_1^n, \ldots, z_i^n \cdot y^n} < 1-\epsilon$

where $R$ is the multiple correlation coefficient;

(iii) there exists $c$ (independent of $n$) such that

$$|y_i^n| \le c \ , \quad |z_{ji}^n| \le c$$

for $n \ge 1$, $1 \le i \le n$ , $1 \le j \le s$ .

Condition (i) is just a normalization of the data, while condition (ii), which can be weakened (see below), seems like a reasonable assumption (e.g.--every house will sell for less than \$20 million, have fewer than 200 baths, etc.). Condition (iii) is necessary to control multiple collinearity.

Let $X^n$ be the subspace generated by $\{z_1^n, \ldots, z_m^n\}$, and $P_{X^n}$ and $P_{Z^n}$ the projections onto $X^n$ and $Z^n$ . By (ii) $X^n$ is m-dimensional, $Z^n$ s-dimensional, and $y^n \notin Z^n$ .

Set $\epsilon^n = y^n - P_{X^n}(y^n)$. For each $\pi \in S_n$ , define $\epsilon_\pi^n$ by $(\epsilon_\pi^n)_i = (\epsilon^n)_{\pi(i)}$ , and let

$$\pi y^n = P_{X^n}(y^n) + \epsilon_\pi^n \ .$$

Let $F_{n\pi}$ be the usual F-statistic for $\pi y^n$ relative to $X^n$ and $Z^n$ :

$$F_{n\pi} = \frac{\frac{1}{s-m} \|P_{X^n}(\pi y^n) - P_{Z^n}(\pi y^n)\|^2}{\frac{1}{n-s} \|\pi y^n - P_{Z^n}(\pi y^n)\|^2} \ .$$

Theorem 2: With notation as above, if conditions (i)-(iii) are satisfied, then the distribution of $\{F_{n\pi}\}_{\pi \in S_n}$ , converges to $\frac{\chi^2_{s-m}}{s-m}$ .

Proof: Fix $n > s+1$, and suppress the index $n$ in the notation introduced above. Since $\pi y = P_X(y) + \epsilon_\pi$ , $P_X(\pi y) = P_X(y) + P_X(\epsilon_\pi)$ and $P_Z(\pi y) = P_X(y) + P_Z(\epsilon_\pi)$ . Thus

$$F_\pi = \frac{\frac{1}{s-m} \|P_X(\epsilon_\pi) - P_Z(\epsilon_\pi)\|^2}{\frac{1}{n-s} \|\epsilon_\pi - P_Z(\epsilon_\pi)\|^2} \ .$$

Orthonormalize $\{z_1, \ldots, z_s, y\}$ by the usual Gram-Schmidt process--obtaining

$\{\tilde{z}_1, \ldots, \tilde{z}_s, \dfrac{\epsilon - P_Z(\epsilon)}{\|\epsilon - P_Z(\epsilon)\|}\}$. By (i), $M_1\{\epsilon\} = M_1\{\tilde{z}_i\} = 1 \quad (1 \le i \le s)$;

and by (ii) and (iii), the coordinates of $\tilde{z}_i$, $\dfrac{\epsilon - P_Z(\epsilon)}{\|\epsilon - P_Z\epsilon\|}$, and $\dfrac{\epsilon}{\|\epsilon\|}$ are

uniformly bounded by $d/\sqrt{n}$ (where $d$ does not depend on $n$).

Now $\|P_X(\epsilon_\pi) - P_Z(\epsilon_\pi)\|^2 = \sum_{i=m+1}^{s} \langle \epsilon_\pi, \tilde{z}_i \rangle^2$, and

$$\|\epsilon_\pi - P_Z(\epsilon_\pi)\|^2 = \|\epsilon\|^2\left(1 - \sum_{i=1}^{s} \frac{\langle \epsilon_\pi, \tilde{z}_i \rangle^2}{\|\epsilon\|^2}\right).$$

Also, $\langle \epsilon_\pi, \tilde{z}_i \rangle = \sum_{j=1}^{n} \epsilon_{\pi(j)} \tilde{z}_{ij}$

$$= \frac{\|\epsilon\|}{\sqrt{n}}\left[\frac{1}{n}\sum_{j=1}^{n}\left(\frac{\sqrt{n}\,\epsilon_{\pi(j)}}{\|\epsilon\|}\right)\left(\sqrt{n}\,\tilde{z}_{ij}\right)\right].$$

By (i) - (iii), the sequences $\left\{\dfrac{\sqrt{n}\,\epsilon_{\pi(j)}}{\|\epsilon\|}\right\}_{1\le j\le n}$ and $\{\sqrt{n}\,\tilde{z}_{ij}\}_{1\le i\le n}$

satisfy the conditions of Theorem 1, so

$$\xi_{\pi i} = \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\left(\frac{\sqrt{n}\,\epsilon_{\pi(j)}}{\|\epsilon\|}\right)\left(\sqrt{n}\,\tilde{z}_{ij}\right)$$

is approximately standard normal under the permutation distribution.

Moreover, for $\lambda_1, \ldots, \lambda_s$ real numbers,

$$\sum_{i=1}^{s} \lambda_i \xi_{\pi i} = \frac{1}{\sqrt{n}}\left[\sum_{j=1}^{n}\left(\frac{\sqrt{n}\,\epsilon_{\pi(j)}}{\|\epsilon\|}\right)\left(\sum_{i=1}^{s}\lambda_i \tilde{z}_{ij} \cdot \frac{\sqrt{n}}{\sqrt{\Sigma\lambda_i^2}}\right)\right] \cdot \sqrt{\Sigma\lambda_i^2}$$

which by Theorem 1 is approximately $N(0, \sum_{i=1}^{s} \lambda_i^2)$ under the permutation

distribution. Hence the permutation distribution of $(\xi_{\pi 1}, \ldots, \xi_{\pi s})$ is

approximately multivariate $N(0,I)$. In particular, $\sum_{i=m+1}^{s} \xi_{\pi i}^2$ is

asymptotically $\chi^2_{s-m}$ and $\sum_{i=1}^{s} \xi_{\pi i}^2$ is asymptotically $\chi^2_s$.

Since

$$F_\pi = \left(\frac{n-s}{n}\right)\left(\frac{1}{s-m}\right)\left(\sum_{i=m+1}^{s} \xi_{\pi i}^2\right)\left(1 - \frac{1}{n}\sum_{i=1}^{s} \xi_{\pi i}^2\right)^{-1} ,$$

the permutation distribution of $F_\pi$ converges to $\frac{\chi^2_{s-m}}{s-m}$ as $n \to \infty$ .

Note: The uniform boundedness condition can be replaced by the following condition. If $\{x_i\}_{n\geq 1}$ is an infinite sequence, and there exists a uniformly bounded infinite sequence $\{x_i'\}_{n\geq 1}$ such that $1/n \sum_{i=1}^{n} (x_i - x_i')^2 \xrightarrow{n\to\infty} 0$ , call $\{x_i\}_{n\geq 1}$ $\ell^2$-u.b. Now suppose there are infinite sequences $\{y_i\}_{i\geq 1}$ , $\{z_i'\}_{i\geq 1}$, $1 \leq j \leq s$ , with $y^n = (y_1, \ldots, y_n)$ and $z_j^n = (z_1^j, \ldots, z_n^j)$. If we replace condition (ii) with

(ii'): $\{y_i\}$ and $\{z_i^j\}_{i\geq 1}$ are $\ell^2$-u.b.

the theorem still holds.

In Section 2, it was claimed that if a population consisting of two types is divided into two groups and the $\chi^2$-statistic for the 2 x 2 table (type x division) is computed, the distribution of this statistic over all possible divisions is approximately $\chi_1^2$ . In Theorem 3, the more general problem in which the population consists of $\ell$ types is considered, and the Theorem gives the stronger result that the distribution of the $\chi^2$-statistic over all divisions into two groups of size $n_1$, and $n_2$ is approximately $\chi^2_{\ell-1}$ . For the asymptotics to apply, of course, neither $n_1$ nor $n_2$ can be too small.

To set up Theorem 3, suppose for each $n$ there is a population of size $n$ consisting of $\ell$ types with $m_i^n$ individuals of type $i$ $(1\leq i\leq \ell)$. A sample of size $n_1^n$ is drawn without replacement from the population, yielding $X_i^n$ individuals of type $i$ . Suppose there exists

nonzero $p$ and $\gamma_i$ $(1 \le i \le \ell)$ such that as $n \to \infty$, $\frac{n_1^n}{n} \to p$ and $\frac{m_i^n}{n} \to \gamma_i$. The distribution of $(X_1^n, \ldots, X_\ell^n)$ is multivariate hypergeometric. Theorem 3 asserts that as $n \to \infty$ with fixed marginal proportions, the distribution of $(X_1^n, \ldots, X_\ell^n)$ is approximately multivariate normal. The $\chi^2_{\ell-1}$ - distribution of the appropriate $\chi^2$-statistic is a corollary of this theorem.

Theorem 3. With the notation of the preceding paragraph, the distribution of $\left(\frac{X_1 - n_p\gamma_1}{\sqrt{np(1-p)}}, \ldots, \frac{X_{\ell-1} - n_p\gamma_{\ell-1}}{\sqrt{np(1-p)}}\right)$ converges to multivariate $N(0,\Sigma)$ where $\Sigma_{ii} = \gamma_i(1-\gamma_i)$ and $\Sigma_{ij} = -\gamma_i\gamma_j$ $(i \neq j)$.

Proof: Fix $n$, and suppress the index $n$ in the notation introduced above. For $\lambda_1, \ldots, \lambda_{\ell-1}$ real, define the list $\{a_i\}_{1 \le i \le n}$ by

$a_1 = \ldots = a_{m_1} = \lambda_1, \ldots, a_{m_1 + \ldots + m_{\ell-2}+1} = \ldots = a_{m_1 + \ldots + m_{\ell-1}} = \lambda_{\ell-1}$,

$a_i = 0$ for $i > m_1 + \ldots + m_{\ell-1}$. Define the list $\{\ell_i\}_{1 \le i \le n}$ by

$\ell_1 = \ldots = \ell_{n_1} = 1$, $\ell_i = 0$ for $i > n_1$.

Set $\bar{\lambda}_n = \lambda_1 \cdot \frac{m_1}{n} + \ldots + \lambda_{\ell-1}\frac{m_{\ell-1}}{n}$ and

$V_n(\lambda) = \lambda_1^2 \frac{m_1}{n} + \ldots + \lambda_{\ell-1}^2 \frac{m_{\ell-1}}{n} - \bar{\lambda}_n^2$. For $1 \le i \le n$ set $x_i = \frac{a_i - \bar{\lambda}_n}{\sqrt{V_n(\lambda)}}$,

so $M_1\{x_i\} = 0$ and $M_2\{x_i\} = 1$.

Similarly, set $y_i = \frac{\ell_i - \frac{n_1}{n}}{\sqrt{\frac{n_1}{n}(1 - \frac{n_1}{n})}}$, so $M_1\{y_i\} = 0$ and $M_2\{y_i\} = 1$.

Since $\{x_i\}$ and $\{y_i\}$ are uniformly bounded (independent of $n$), Theorem 1 may be applied to yield:

$$\left\{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_i y_{\pi(i)}\right\}_{\pi \in S_n} \sim N(0,1) .$$

Since $\sum_{i=1}^{n} x_i y_{\pi(i)} = \sum_{i=1}^{\ell-1} \lambda_i X_i$ and $\bar{\lambda}_n \to \sum_{i=1}^{\ell-1} \lambda_i \gamma_i \overset{\text{def}}{=} \bar{\lambda}$ and

$V_n(\lambda) \to \sum_{i=1}^{\ell-1} \lambda_i^2 \gamma_i - \bar{\lambda}^2 \overset{\text{def}}{=} V(\lambda)$ , we have

$$\sum_{i=1}^{\ell-1} \frac{\lambda_i}{V(\lambda)} \left(\frac{X_i^n - n_1 \gamma_i}{\sqrt{np(1-p)}}\right) \to N(0,1) \text{ as } N \to \infty .$$

Since this holds for each $(\lambda_1, \ldots, \lambda_{\ell-1}) \in R^{\ell-1}$ , the Theorem follows.

Note: This theorem can be generalized to the case in which a population consisting of $\ell$ types is divided into r groups. The proof extends the proof above by an argument which depends on obtaining the joint distribution of the number of each type drawn in $(r-1)$ successive samples without replacement of fixed proportions of the population. This proof can be based on a generalization of Theorem 1 which will appear in a later paper by the present authors.

## APPENDIX

Example 2: Data from Narula and Wellington (1977)

| Case | Variable Number | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | Y |
| 1 | 4.9176 | 1.0 | 3.4720 | 0.9980 | 1.0 | 7 | 4 | 42 | 3 | 1 | 0 | 25.9 |
| 2 | 5.0208 | 1.0 | 3.5310 | 1.5000 | 2.0 | 7 | 4 | 62 | 1 | 1 | 0 | 29.5 |
| 3 | 4.5429 | 1.0 | 2.2750 | 1.1750 | 1.0 | 6 | 3 | 40 | 2 | 1 | 0 | 27.9 |
| 4 | 4.5573 | 1.0 | 4.0500 | 1.2320 | 1.0 | 6 | 3 | 54 | 4 | 1 | 0 | 25.9 |
| 5 | 5.0597 | 1.0 | 4.4550 | 1.1210 | 1.0 | 6 | 3 | 42 | 3 | 1 | 0 | 29.9 |
| 6 | 3.8910 | 1.0 | 4.4550 | 0.9880 | 1.0 | 6 | 3 | 56 | 2 | 1 | 0 | 29.9 |
| 7 | 5.8980 | 1.0 | 5.8500 | 1.2400 | 1.0 | 7 | 3 | 51 | 2 | 1 | 1 | 30.9 |
| 8 | 5.6039 | 1.0 | 9.5200 | 1.5010 | 0.0 | 6 | 3 | 32 | 1 | 1 | 0 | 28.9 |
| 9 | 15.4202 | 2.5 | 9.8000 | 3.4200 | 2.0 | 10 | 5 | 42 | 2 | 1 | 1 | 84.9 |
| 10 | 14.4598 | 2.5 | 12.800 | 3.0000 | 2.0 | 9 | 5 | 14 | 4 | 1 | 1 | 82.9 |
| 11 | 5.8282 | 1.0 | 6.4350 | 1.2250 | 2.0 | 6 | 3 | 32 | 1 | 1 | 0 | 35.9 |
| 12 | 5.3003 | 1.0 | 4.9883 | 1.5520 | 1.0 | 6 | 3 | 30 | 1 | 2 | 0 | 31.5 |
| 13 | 6.2712 | 1.0 | 5.5200 | 0.9750 | 1.0 | 5 | 2 | 30 | 1 | 2 | 0 | 31.0 |
| 14 | 5.9592 | 1.0 | 6.6660 | 1.1210 | 2.0 | 6 | 3 | 32 | 2 | 1 | 0 | 30.9 |
| 15 | 5.0500 | 1.0 | 5.0000 | 1.0200 | 0.0 | 5 | 2 | 46 | 4 | 1 | 1 | 30.0 |
| 16 | 5.6039 | 1.0 | 9.5200 | 1.5010 | 0.0 | 6 | 3 | 32 | 1 | 1 | 0 | 28.9 |
| 17 | 8.2464 | 1.5 | 5.1500 | 1.6640 | 2.0 | 8 | 4 | 50 | 4 | 1 | 0 | 36.9 |
| 18 | 6.6969 | 1.5 | 6.9020 | 1.4880 | 1.5 | 7 | 3 | 22 | 1 | 1 | 1 | 41.9 |
| 19 | 7.7841 | 1.5 | 7.1020 | 1.3760 | 1.0 | 6 | 3 | 17 | 2 | 1 | 0 | 40.5 |
| 20 | 9.0384 | 1.0 | 7.8000 | 1.5000 | 1.5 | 7 | 3 | 23 | 3 | 3 | 0 | 43.9 |
| 21 | 5.9894 | 1.0 | 5.5200 | 1.2560 | 2.0 | 6 | 3 | 40 | 4 | 1 | 1 | 37.5 |
| 22 | 7.5422 | 1.5 | 4.0000 | 1.6900 | 1.0 | 6 | 3 | 22 | 1 | 1 | 0 | 37.9 |
| 23 | 8.7951 | 1.5 | 9.8900 | 1.8200 | 2.0 | 8 | 4 | 50 | 1 | 1 | 1 | 44.5 |
| 24 | 6.0931 | 1.5 | 6.7265 | 1.6520 | 1.0 | 6 | 3 | 44 | 4 | 1 | 0 | 37.9 |
| 25 | 8.3607 | 1.5 | 9.1500 | 1.7770 | 2.0 | 8 | 4 | 48 | 1 | 1 | 1 | 38.9 |
| 26 | 8.1400 | 1.0 | 8.0000 | 1.5040 | 2.0 | 7 | 3 | 3 | 1 | 3 | 0 | 36.9 |
| 27 | 9.1416 | 1.5 | 7.3262 | 1.8310 | 1.5 | 8 | 4 | 31 | 4 | 1 | 0 | 45.8 |
| 28 | 12.0000 | 1.5 | 5.0000 | 1.2000 | 2.0 | 6 | 3 | 30 | 3 | 1 | 1 | 41.0 |

$X_1$ = Taxes (100's of dollars)

$X_2$ = No. of baths

$X_3$ = Lot size/1000 $ft^2$

$X_4$ = Living space/1000 $ft^2$

$X_5$ = No. of garages

$X_6$ = No. of rooms

$X_7$ = No. of bedrooms

$X_8$ = Age of house

$X_9$ = Construction type (coded 1,2,3,4)

$X_{10}$ = Style (coded 1,2,3)

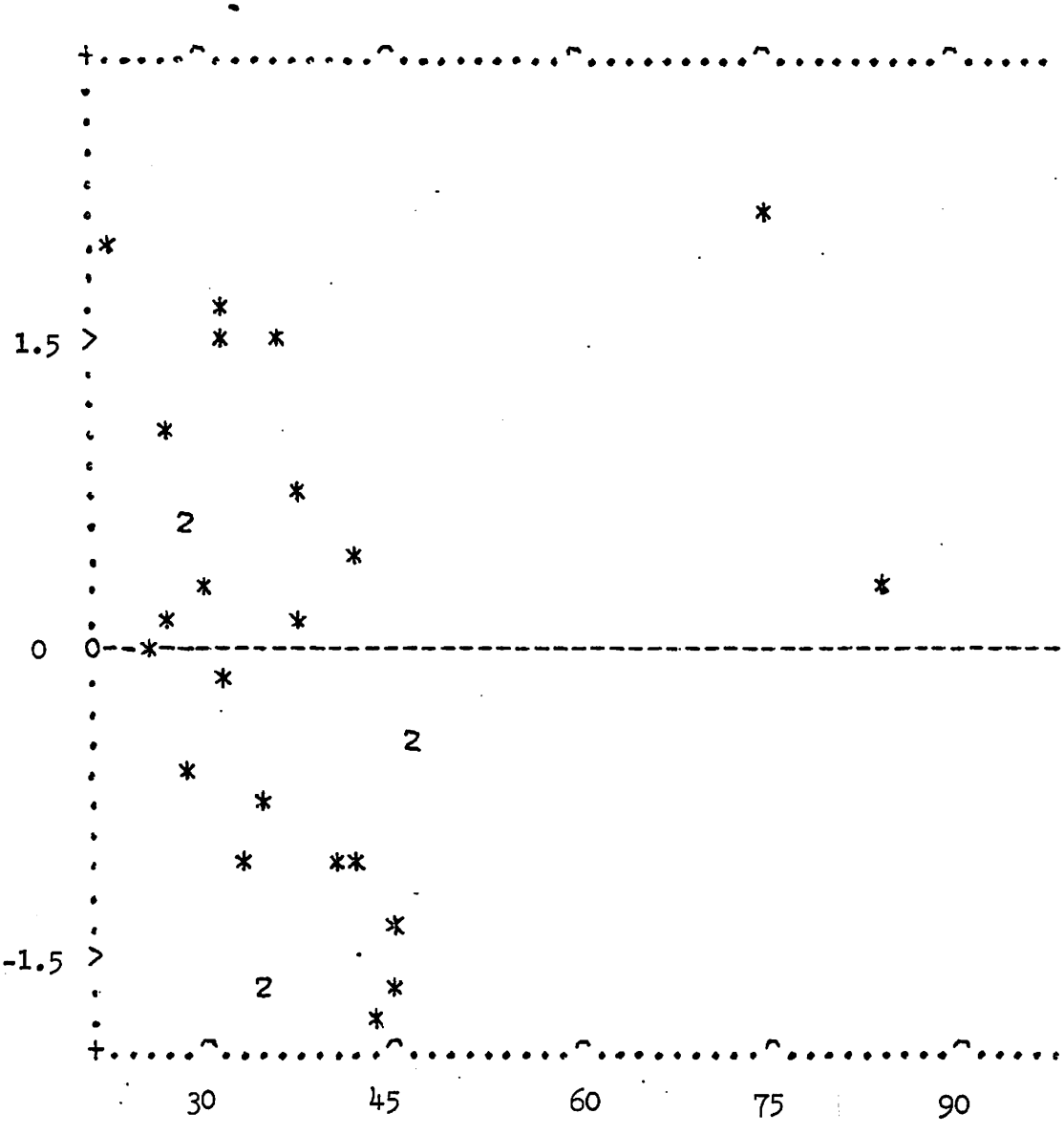
$X_{11}$ = No. of fireplaces

Y = Sale price (1000's of dollars)

RESIDUAL PLOTS

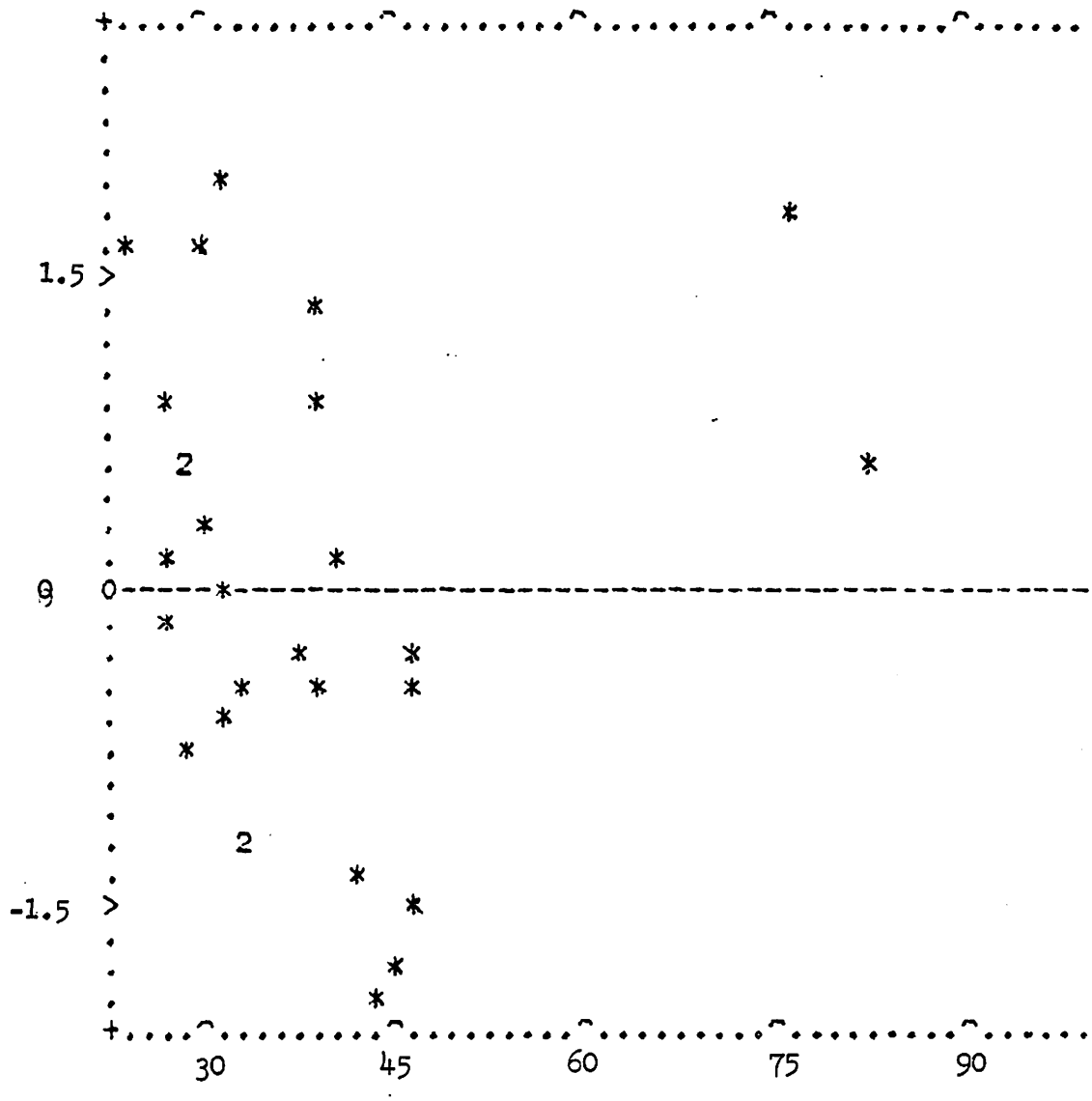
X: Fitted values for the regression
of selling price on taxes and living space.

Y: Residuals



X: Fitted values for the regression
on taxes, living space, and number of baths

Y: Residuals



X: Number of baths

Y: Residuals for the regression of selling price on taxes and living space