# THE INTERFACE BETWEEN STATISTICAL METHODOLOGY AND STATISTICAL PRACTICE

Gary G. Koch, University of North Carolina, Chapel Hill

## 1. Introduction

This paper is concerned with philosophical issues which require consideration whenever statistical methodology is applied to data. For this purpose, attention is focused on certain essential questions which statisticians must address for their efforts to be more meaningful than misleading. These include:

1. distinction between study population and target population,
2. distinction between variables under study and concepts which they are operationally assumed to represent,
3. role of technical assumptions pertaining to research design, existing state of knowledge, and statistical framework in which study objectives are formulated.

These and other aspects of statistical practice share "context" as a common theme. Here, "context" represents a perspective for evaluating the validity of the use of a particular statistical method through the relationship of the interpretation of its results to the specific nature of individual applications. Further clarification of this point of view is given for such topics as variable scaling, variable selection, and model building as applied to observational data, experimental data, and population sample survey data. For this purpose, an outline format discussion is given for two examples.

## ACKNOWLEDGMENTS

This research was in part supported through a Joint Statistical Agreement with Burroughs Wellcome Company. The author would like to thank Jean Harrison, Jean McKinney and Pat Peek for their conscientious typing of this manuscript.

## REFERENCES

Bishop, Y.M.M., Fienberg, S.E., Holland, P.W. Discrete Multivariate Analysis (M.I.T. Press, 1975).

Grizzle, J.E., Starmer, C.F., Koch, G.G. (1969). Analysis of categorical data by linear models. Biometrics 25, 489-504.

Higgins, J.E., Koch, G.G. (1977). Variable selection and generalized chi-square analysis of categorical data applied to a large cross-sectional occupational health survey. International Statistical Review 45, 51-62.

Koch, G.G., Freeman, D.H.,Jr., Freeman, J.L. (1975). Strategies in the multivariate analysis of data from complex surveys. International Statistical Review 43, 59-78.

Koch, G.G., Freeman, J.L., Lehnen, R.G. (1976). A general methodology for the analysis of ranked policy preference data. International Statistical Review 44, 1-28.

Koch, G.G., et al. (1976). The asymptotic covariance structure of estimated parameters from contingency table log-linear models. Proceedings of the 9th International Biometric Conference, 317-336.

Koch, G.G., et al. (1977). A general methodology for the analysis of experiments with repeated measurement of categorical data. Biometrics 33, 133-158.

Landis, J.R., et al. (1977) Parcat: a computer program for testing average partial association in three-way contingency tables. 1977 Proceedings of ASA Statistical Computing Section.

---

Example 1: Observational Data from a Case History Record System

a. Source: Clarke, S.H. and Koch, G.G. (1976). The influence of income and other factors on whether criminal defendants go to prison, Law and Society Review Volume 11, pp. 57-92.

b. Subject Matter and Objectives: To study historically a sample of persons arrested for certain types of burglary and larceny and to evaluate the extent to which an active prison sentence outcome was related to variables pertaining to the defendant's demographic status, specific type of offense, prior arrest record, etc.

c. Sample Design: All persons who were arrested for burglary, breaking and entering, and larceny (excluding automobile thefts and thefts involving less than $5.00) in Mecklenburg County, North Carolina with prosecutions begun during 1971. There were 798 such persons and all of them are included in the sample. Thus, the sample here corresponds to a total population.

d. Target Population:
   i. Local inferences: The population of interest is the actual sampled population which is restricted in time and place to 1971 and Mecklenburg County, N.C.
   ii. Extended inferences: The super-population of all persons who have been, are, or eventually will be arrested for burglary, breaking and entering, and larceny regardless of time and place from which the sampled population can be hypothetically regarded as a stratified simple random

sample with the strata being the cells of the multi-way cross-classification of those demographic, offense type, prior arrest record, etc. variables which have a statistically important relationship with whether a defendant receives an active prison sentence or not. In this regard, it should be noted that such a hypothetical super-population may not exist in which case any extended inferences are meaningless from a practical point of view. Nevertheless, an awareness of the existence of a context where they may be appropriate is still of interest.

e. Variables Under Study:
   i. Prison sentence status (Yes, No)
   ii. Type of offense charged (Non-residential burglary: NRB, Residential burglary: RB, Felonious and misdemeanor larceny: LARC)
   iii. Prior arrests (None: 0, One or more: 1+)
   iv. Arrest promptness (Same day: S, Later day: L). This variable is regarded as a measure of strength of evidence since it seems reasonable to assume that arrests which occurred very soon after the offense would tend to be based on more specific evidence (as opposed to circumstantial evidence) than those which occurred later.
   v. Median income of census tract of residence (Less than $7,000: L, At least $7,000 or suburban residents with unclassified income in terms of this definition: H). This variable is regarded as a general measure of socio-economic status as opposed to specific earnings.
   vi. Other variables which were considered included age, race, sex, and employment. However, after (ii)-(v) were taken into account, these other variables did not have a statistically important relationship with prison sentence status.

f. Data Display: The data corresponding to the multiway cross-classification of offense x prior arrests x arrest promptness x income x prison sentence status are summarized in contingency table format in Table 1.

TABLE 1

BURGLARY-LARCENY DATA: MULTI-WAY CROSS-CLASSIFICATION OF OFFENSE x PRIOR ARRESTS x ARREST PROMPTNESS x INCOME x DEFENDANT'S PRISON SENTENCE STATUS

| Offense | Prior Arrests | Arrest Promptness | Income | Defendant's Prison Status Yes | Defendant's Prison Status No | Observed Prison Proportion | Est. s.e. | Statistical Model $X$ | Model Predicted Prison Proportion | Est. s.e. |
|---|---|---|---|---|---|---|---|---|---|---|
| NRB | 1+ | S | L | 15 | 14 | 0.517 | 0.093 | 1 2 0 | 0.537 | 0.050 |
| NRB | 1+ | S | H | 4 | 11 | 0.267 | 0.114 | 1 1 0 | 0.304 | 0.025 |
| NRB | 1+ | L | L | 12 | 22 | 0.353 | 0.082 | 1 1 0 | 0.304 | 0.025 |
| NRB | 1+ | L | H | 11 | 20 | 0.355 | 0.086 | 1 1 0 | 0.304 | 0.025 |
| NRB | 0 | S | L | 7 | 5 | 0.583 | 0.142 | 1 2 0 | 0.537 | 0.050 |
| NRB | 0 | S | H | 3 | 8 | 0.273 | 0.134 | 1 1 0 | 0.304 | 0.025 |
| NRB | 0 | L | L | 6 | 12 | 0.333 | 0.111 | 1 1 0 | 0.304 | 0.025 |
| NRB | 0 | L | H | 1 | 13 | 0.071 | 0.069 | 1 0 0 | 0.072 | 0.013 |
| RB | 1+ | S | L | 10 | 20 | 0.333 | 0.086 | 1 1 0 | 0.304 | 0.025 |
| RB | 1+ | S | H | 1 | 4 | 0.200 | 0.179 | 1 1 0 | 0.304 | 0.025 |
| RB | 1+ | L | L | 15 | 36 | 0.294 | 0.064 | 1 1 0 | 0.304 | 0.025 |
| RB | 1+ | L | H | 4 | 32 | 0.111 | 0.052 | 1 0 0 | 0.072 | 0.013 |
| RB | 0 | S | L | 2 | 8 | 0.200 | 0.126 | 1 1 0 | 0.304 | 0.025 |
| RB | 0 | S | H | 1 | 4 | 0.200 | 0.179 | 1 1 0 | 0.304 | 0.025 |
| RB | 0 | L | L | 1 | 17 | 0.055 | 0.054 | 1 0 0 | 0.072 | 0.013 |
| RB | 0 | L | H | 1 | 19 | 0.050 | 0.049 | 1 0 0 | 0.072 | 0.013 |
| LARC | 1+ | S | L | 15 | 51 | 0.227 | 0.052 | 1 0 1 | 0.193 | 0.032 |
| LARC | 1+ | S | H | 5 | 38 | 0.116 | 0.049 | 1 0 0 | 0.072 | 0.013 |
| LARC | 1+ | L | L | 14 | 68 | 0.171 | 0.042 | 1 0 1 | 0.193 | 0.032 |
| LARC | 1+ | L | H | 3 | 53 | 0.054 | 0.030 | 1 0 0 | 0.072 | 0.013 |
| LARC | 0 | S | L | 2 | 24 | 0.077 | 0.052 | 1 0 0 | 0.072 | 0.013 |
| LARC | 0 | S | H | 6 | 66 | 0.083 | 0.033 | 1 0 0 | 0.072 | 0.013 |
| LARC | 0 | L | L | 5 | 53 | 0.086 | 0.037 | 1 0 0 | 0.072 | 0.013 |
| LARC | 0 | L | H | 3 | 53 | 0.054 | 0.030 | 1 0 0 | 0.072 | 0.013 |

g. Data Analysis Strategies:

i. Local inferences. The basic framework is the multiple hypergeometric Model 0 in Appendix 1 with respect to which the hypothesis of randomness is being tested within two-way tables with fixed marginals and within sets of two-way tables with fixed margins. Of course, in a strict sense, all of the frequency counts are fixed constants (as opposed to random variables) because of the historical nature of the data. On the other hand, one can argue that there is still interest in the hypothetical question of whether or not the observed distribution of prison sentence status is at random with respect to each of the arrest description variables under study (for both the entire population as well as for sub-populations which are based on the other variables which are not being tested).

ii. Extended inferences. If the results of the local inference analysis seem plausible with respect to existing knowledge or theory for the substantive subject matter field to which the conclusions of the study are to be directed, then it may be realistic to assume the existence of a potential super-population to which such conclusions can be extended. In this case, the basic framework for analysis is the product multinomial Model 1 in Appendix 2. Thus, the respective proportions of defendants receiving prison sentences are random variables, and the principal objective of analysis is the characterization of the variation among them through the fitting of regression models and the testing of various hypotheses involving their parameters.

h. Results

i. Local inferences. Pearson chi-square statistics $Q_P$ for testing the significance of the relationship between prison sentence status and the arrest descriptor variables are shown below.

| Prison x Offense | Prison x Prior Arrests | Prison x Arrest Promptness | Prison x Income |
|---|---|---|---|
| $Q_P$(D.F.=2) = 48.35 | $Q_P$(D.F.=1) = 15.23 | $Q_P$(D.F.=1) = 4.43 | $Q_P$(D.F.=1) = 19.45 |

Thus, offense, prior arrests, and income are significantly ($\alpha$=0.01) related to prison sentence status in a strong sense (either with or without adjustment for multiple comparisons via Bonferroni inequality methods). However, the relationship between arrest promptness and prison sentence is only significant ($\alpha$=0.05) in the weak sense where multiple comparison issues are ignored. Thus, caution should be exercised with respect to the nature of conclusions concerning this relationship (unless it was the one of primary interest in which case the other relationships would only be investigated from a descriptive as opposed to an inferential point of view).

Since the first order relationship of prison status to some of the arrest descriptor variables may be strongly influenced by the relationship of such variables to each other, partial association tests become of interest. For example, if the population is partitioned into three sets corresponding to offense type, to what extent is prison sentence status significantly related to income within these respective sets (taken together as a whole). A valid test statistic for this hypothesis (if the sample sizes within each set are sufficiently large) is the sum of the Pearson chi-square statistics for prison sentence status vs. income for the three offense types. Since $Q_{TP}$(D.F.=3) = 17.70, this partial association relationship is significant ($\alpha$=0.01) with multiple comparisons issues being ignored since this test is typically in a philosophically different class than the ones described previously (i.e., either this type or the previous type or some third type may be regarded as the tests of primary interest from an inferential point of view but not all simultaneously since the spirit underlying the use of multiple tests here is the descriptive demonstration of support for a conclusion from several different points of view as opposed to the search for "significance" in the midst of randomness). If this type of analysis is continued further, the partial association between prior arrest history and prison sentence status after adjustment for (the joint partition of) offense type and income is considered. Here, however, the Cochran-Mantel-Haenszel statistic for which D.F.=1 is used in order to direct statistical power at average partial association alternatives (i.e., the extent to which the direction of the relationship between prior arrest history and prison sentence status is consistent across the six offense type x income sub-populations even though some of their respective magnitudes may be small). Since $Q_{CMH}$(D.F.=1) = 8.00, this partial association is significant ($\alpha$=0.01). Finally, the partial association of arrest promptness with prison sentence status after adjustment for (the joint partition of) offense type, income, and prior arrest history is significant ($\alpha$=0.05) with $Q_{CMH}$ = 5.42.

In summary, several different types of hypotheses can be investigated for the purpose of local inferences about certain types of relationships among variables in populations of observational data. However, since the randomness in the data is only induced through the consideration of hypotheses, other types of statistical analysis like the estimation of standard errors for observed proportions, measures of association, etc. and the construction of confidence intervals cannot be undertaken in this framework because the data do indeed correspond to a population rather than to a sample from a population.

ii. Extended inferences. Here, the weighted least squares methods described in Grizzle, Starmer, and Koch [1969] are used to investigate the nature of the variation among the probabilities of defendants receiving prison sentences for the respective super-sub-populations corresponding to the (offense type x prior arrest history x arrest promptness x income) cross-classification. Since no prior information is available concerning the specific structure of a statistical

model for characterizing such variation, the complete contingency table is partitioned into six modules on the basis of offense type and prior arrest history, the two most important variables from a substantive point of view. Separate analyses are then undertaken within each of these modules in a manner which is primarily oriented toward their individual features but also attempts to reflect descriptively any consistency among them. As a result, the following models are found to be appropriate for the six (offense type x prior arrest history) modules

Non-residential Burglary
One or more prior arrests

$$\underset{\sim}{X}\,\underset{\sim}{b} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0.335 \\ 0.182 \end{bmatrix}$$

Model Q(D.F.=1) = 2.91
Residual Q(D.F.=2) = 0.46

Residential Burglary
One or more prior arrests

$$\underset{\sim}{X}\,\underset{\sim}{b} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0.111 \\ 0.189 \end{bmatrix}$$

Model Q(D.F.=1) = 6.89
Residual Q(D.F.=2) = 0.47

Larceny
One or more prior arrests

$$\underset{\sim}{X}\,\underset{\sim}{b} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0.071 \\ 0.122 \end{bmatrix}$$

Model Q(D.F.=1) = 8.76
Residual Q(D.F.=2) = 1.92

---

Non-residential Burglary
No prior arrests

$$\underset{\sim}{X}\,\underset{\sim}{b} = \begin{bmatrix} 1 & 2 \\ 1 & 1 \\ 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0.068 \\ 0.251 \end{bmatrix}$$

Model Q(D.F.=1) = 12.22
Residual Q(D.F.=2) = 0.15

Residential Burglary
No prior arrests

$$\underset{\sim}{X}\,\underset{\sim}{b} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0.052 \\ 0.148 \end{bmatrix}$$

Model Q(D.F.=1) = 1.82
Residual Q(D.F.=2) = 0.01

Larceny
No prior arrests

$$\underset{\sim}{X}\,\underset{\sim}{b} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \begin{bmatrix} 0.073 \end{bmatrix}$$

Residual Q(D.F.=3) = 0.65

After noting certain similarities among the predicted values across the six modules, the six distinct models are then synthesized together to form the overall model shown in Table 1 by methods analogous to those used in Koch, Freeman and Lehnen [1976] and Higgins and Koch [1977]. This model provides a relatively complete characterization of the variation among the proportions of defendants receiving prison sentences since

Model Q(D.F.=2) = 81.70 Residual Q(D.F.=21) = 6.01

Thus, the corresponding predicted proportions in Table 1 which are based on it represent a useful descriptive summary of the relationship between prison sentence status and offense type, prior arrest history, arrest promptness, and income in the hypothetical super-population from which the data are presumed to have arisen and to which any inferences are directed.

As a final comment, it should be noted that the structure of this model suggests the presence of substantial interaction among the respective arrest descriptor variables since the nature of the relationships within modules varies across modules. This aspect of the analysis may seem troublesome because the significance of such interaction has not been demonstrated. However, for this investigation tests for such interaction are not of direct interest because they pertain more directly to model reduction strategies than to substantively important hypotheses. For this reason, instead of reflecting a type of conclusion in the usual sense, interaction here corresponds to a concept in terms of which conclusions are qualified; i.e., the structure of the model for one module is not necessarily forced on the others unless the relationships within them are clearly compatible, when considered separately in their own right rather than with respect to the results of a statistical testing procedure, which sometimes are relatively weak (because they are directed at residuals of the models).

i. Conclusions

i. Local inferences. There do exist statistically significant relationships between prison sentence status and offense type, prior arrest history, arrest promptness, and income in the historical sample population of arrests for burglary, breaking and entering, and larceny in Mecklenburg County, North Carolina with prosecutions begun during 1971.

ii. Extended inferences. Since the sampled population in this application is sufficiently narrow to be of very limited interest in its own right, it is potentially desirable to argue that its local inferences can be extrapolated to some larger target super-population, even though it may not be possible to specify directly its location in time and place. This point of view is supported by the fact that the results of analysis are plausible with respect to existing knowledge in the criminal justice area (i.e., those relationships which are found to be statistically important are also, for the most part, substantively meaningful in both direction and magnitude). Thus, the structure of the fitted model $\underset{\sim}{X}$ in Table 1 and its corresponding predicted values for the data from this investigation are considered to be of general descriptive interest with respect to prison sentence outcome for burglary and larceny arrests, subject to the fundamental caution that any conclusions which are based on them should be regarded as

inherently tentative until it receives further support by similarly oriented studies for other locations and time periods.

j. Other sources of examples for observational data from case history records include:
   i. Analyses of highway safety injury data from motor vehicle accident populations which are defined in terms of record files for specific states and time periods;
   ii. Analyses of medical or dental outcome data for patient populations which are defined in terms of record files for specific clinics and time periods;
   iii. Analyses of product performance or safety history for consumer populations which are defined in terms of record files for specific distributors and time periods.

k. Summary statement concerning methodological issues: The most critical consideration underlying the interpretation of the analysis of this type of data is the relevance of the sampled population to the target population. This issue applies equally strongly for when the sample under study is a sample of a case history record system, as opposed to a total population like the one discussed here. Data quality and certain technical aspects of their statistical behavior are also important, but can often be assumed to satisfy the required conditions, since the scope of such analyses is either hypothetical or restricted to the descriptive summary of an isolated population of operationally prepared records (as opposed to measured phenomena). In other words, observational data from case history records basically stand on their own within whatever specific framework evolves. Thus, the principal question of interest is whether or not their analysis can be interpreted more broadly.

Example 2: Experimental Design Data

a. Source: Grizzle, J. E., Starmer, C. F. and Koch, G. G. (1969). Analysis of categorical data by linear models, Biometrics, Volume 25, pp. 489-504.

b. Subject Matter and Objectives: To investigate the relationship between the severity of the "dumping syndrome," an undesirable sequela of surgery for duodenal ulcer, and the nature and extent of surgery for four different types of operations which involve the removal of different amounts of the stomach.

c. Experimental Design: A multi-clinic randomized clinical trial involving suitably eligible patients who were treated in four participating hospitals during approximately 1966-1968.

d. Target Population:
   i. Local inferences: The population of interest is the actual study population as defined by the protocol inclusion criteria, the 1966-1968 time period, and the four hospitals.
   ii. Extended inferences: The super-population of all persons who have been, are, or eventually will be treated by one of the four operations regardless of time and place, from which the study population can be hypothetically regarded as a stratified simple random sample with the strata being the cells of the multi-way cross-classification of hospital, operation type, and any relevant demographic or patient diagnostic variables which have a statistically important relationship with the severity of the dumping syndrome which a patient experiences.

e. Variables Under Study:
   i. Dumping syndrome severity (None: N, Slight: S, Moderate: M)
   ii. Operation (Drainage and vagotomy: 0, Antrectomy (25% resection) and vagotomy: 1, Hemigastrectomy (50% resection) and vagotomy: 2, and 75% resection: 3)
   iii. Hospital (Hospital 1, Hospital 2, Hospital 3, Hospital 4)

f. Data Display: The data corresponding to the multiway cross-classification of hospital x operation x dumping syndrome severity are summarized in contingency table format in Table 2.

g. Data Analysis Strategies:
   i. Local inferences. The basic framework is the multiple hypergeometric Model 0 in Appendix 1 with respect to which the hypothesis of randomness is being tested for the relationship between operation and dumping syndrome severity in the set of four two-way tables corresponding to the respective hospitals. Here, the marginal distributions of operation type are regarded as fixed in principle by the nature of the experimental design (but may actually be subject to some inherent random variability because of possible protocol violations, missing data, etc.). In addition, the marginal distribution of the dumping syndrome severity is regarded as fixed under the null hypothesis (because it implies that the dumping syndrome severity for each separate patient is not affected by the operation that is experienced and thus, its distribution remains the same for all realizations of the treatment randomization process). Otherwise, since the dumping syndrome severity data are ordinally scaled, the types of alternatives which are of primary interest are location shifts which are indicative of the extent to which the dumping syndrome tends to be more severe for certain operations than others. It is also of interest to investigate the extent to which these location shifts are related to the ordinal scaling of the operations with respect to the amount of stomach removed.
   ii. Extended inferences. If the local inference results indicate a significant difference which is considered to be generalizable to some larger population, then it becomes realistic to analyze the data in terms of the product multinomial Model 1 in Appendix 2. In this regard, the

TABLE 2

DUMPING SYNDROME DATA: MULTI-WAY CROSS-CLASSIFICATION OF HOSPITAL × OPERATION × DUMPING SYNDROME SEVERITY

| Hospital | Operation | Dumping Syndrome Severity N | S | M | Normalized Uniform Average Score | Est. s.e. | Statistical Model $\underset{\sim}{X}$ | Model Predicted Average Score | Est. s.e. |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 23 | 7 | 2 | 0.17 | 0.05 | 1 0 | 0.20 | 0.03 |
| 1 | 1 | 23 | 10 | 5 | 0.26 | 0.06 | 1 1 | 0.24 | 0.02 |
| 1 | 2 | 20 | 13 | 5 | 0.30 | 0.06 | 1 2 | 0.28 | 0.02 |
| 1 | 3 | 24 | 10 | 6 | 0.28 | 0.06 | 1 3 | 0.33 | 0.03 |
| 2 | 0 | 18 | 6 | 1 | 0.16 | 0.05 | 1 0 | 0.20 | 0.03 |
| 2 | 1 | 18 | 6 | 2 | 0.19 | 0.06 | 1 1 | 0.24 | 0.02 |
| 2 | 2 | 13 | 13 | 2 | 0.30 | 0.06 | 1 2 | 0.28 | 0.02 |
| 2 | 3 | 9 | 15 | 2 | 0.36 | 0.06 | 1 3 | 0.33 | 0.03 |
| 3 | 0 | 8 | 6 | 3 | 0.36 | 0.09 | 1 0 | 0.20 | 0.03 |
| 3 | 1 | 12 | 4 | 4 | 0.30 | 0.09 | 1 1 | 0.24 | 0.02 |
| 3 | 2 | 11 | 6 | 2 | 0.26 | 0.08 | 1 2 | 0.28 | 0.02 |
| 3 | 3 | 7 | 7 | 4 | 0.42 | 0.09 | 1 3 | 0.33 | 0.03 |
| 4 | 0 | 12 | 9 | 1 | 0.25 | 0.06 | 1 0 | 0.20 | 0.03 |
| 4 | 1 | 15 | 3 | 2 | 0.18 | 0.07 | 1 1 | 0.24 | 0.02 |
| 4 | 2 | 14 | 8 | 3 | 0.28 | 0.07 | 1 2 | 0.28 | 0.02 |
| 4 | 3 | 13 | 6 | 4 | 0.30 | 0.08 | 1 3 | 0.33 | 0.03 |

variation among certain mean scores is investigated through the fitting of regression models and the testing of various hypotheses involving their parameters.

h. Results

i. Local inferences. The Cochran-Mantel-Haenszel statistic is used to test the significance of the partial association between the operation type and the dumping syndrome severity after adjustment for (the partition of) hospital. Moreover, to target statistical power at order partial association alternatives (i.e., the extent to which the probability of more severe dumping syndrome outcomes tends to increase (or decrease) with the extent of the operation in terms of larger amounts of stomach removed), the correlation mode for this statistic with D.F.=1 is used. In this regard, two types of scores are potentially appropriate. The first is ridits (or equivalently rank scores) which provides a partial Spearman rank correlation analysis of the data. The principal advantage of this approach is that it provides a framework in which the potentially difficult question of variable scaling can be avoided. Its disadvantage is that its results do not necessarily have a straightforward interpretation with respect to such scales since the analysis proceeds in terms of an index. Alternatively, uniform (or normalized uniform) mean scores can be used. Here, the functions $(p_S + p_M)$ and $p_M$ are regarded from a substantive point of view as equally important measures of dumping syndrome severity for the purpose of assessing variation among the operations (on a within hospital basis). Thus, if there is no variation among the operations, there is no variation with respect to each of these measures and hence no variation with respect to their sum $(0p_N + p_S + 2p_M)$. Alternatively, if there is variation with respect to their sum, there must be some variation with respect to either $p_S + p_M$ and/or $p_M$. Similarly, if $F_0$, $F_1$, $F_2$, $F_3$ are measures of dumping syndrome severity for operations 0, 1, 2, 3 respectively, then the ordered pairwise differences $(F_1-F_0)$, $(F_2-F_0)$, $(F_3-F_0)$, $(F_2-F_1)$, $(F_3-F_1)$, $(F_3-F_2)$ will all be expected to be null if there is no variation among the four operations. Thus, their sum $G = (-3F_0 - F_1 + F_2 + 3F_3)$ is also expected to be null. However, if this sum is concluded to be non-null, then there must be some variation among the four operations. Otherwise, it should be noted that the function G is constructed to compound pairwise differences which are arranged to reinforce one another if indeed the probability of more severe dumping syndrome outcomes does tend to increase (or decrease) with the extent of the operation. Thus, uniform scores can be used for both dumping syndrome severity and operation on the basis of statistical power arguments with respect to order association alternatives. In addition, although the scaling which they induce on the categories for these variables may not necessarily have a meaningful substantive interpretation, they do nevertheless provide a quantitative framework which can be used for descriptive statistical purposes.

For this example, the normalized uniform scores (0, 0.5, 1) will be used for the dumping syndrome so that "moderate" is regarded as the principal response level of interest and "slight" is interpreted as half-way between in the sense that two people with "slight" are considered to be equivalent for comparison purposes to one person with "none" and one person with "moderate." Since the operations are naturally scaled with respect to the amount of stomach removed, the scaling 0, 1, 2, 3 does not require any further explanation.

Finally $Q_{CMH}(D.F.=1) = 6.34$ for uniform scores and $Q_{CMH}(D.F.=1) = 6.92$ for ridit scores. The former is significant at ($\alpha=0.05$) and the latter is significant at ($\alpha=0.01$). Thus, both indicate a significant relationship between dumping syndrome severity and the extent of the operation. Moreover, it should be noted that the focus of these statistics on order association alternatives at the beginning is critical because the overall Cochran-Mantel-Haenszel statistic $Q_{CMH}(D.F.=6) = 10.60$ is not significant ($\alpha=0.10$).

In summary, the partial association between dumping syndrome severity and extent of operation can be investigated for the purpose of local inferences with respect to the set of patients defined by the protocol inclusion criteria, the 1966-1968 time period, and the four hospitals. For this purpose, the only assumption required is the validity of the randomization process by which patients were assigned to operations. However, once the hypothesis of randomness is rejected, the hypergeometric Model 0 in Appendix 1 is no longer applicable. In addition, since such rejection implies the existence of a significant relationship in a local inference sense, it then becomes of interest to characterize descriptively its nature in terms of fitted regression models for extended inference purposes with respect to some larger super-population.

ii. Extended inferences. As with Example 1, the weighted least squares methods in Grizzle, Starmer, and Koch are used to investigate the nature of the variation of the distribution of dumping syndrome severity for the respective super-sub-populations corresponding to the operation x hospital cross-classification. More specifically, attention will be focused on the mean score function $F = (0.5p_S + p_M)$ because of its sensitivity to location shifts and its compatibility with certain asymptotic (central limit theory) assumptions as discussed in Koch et al. [1977]. Secondly, since the participating hospitals all followed the same basic protocol with respect to the inclusion of eligible patients in the study, the conduct of the four operations, and the evaluation of patient response, it is reasonable to assume a priori that the variation among the mean score functions $F_{hi}$ (where h = 1, 2, 3, 4 indexes hospitals and i = 0, 1, 2, 3 indexes operations) can be characterized in terms of an additive model with respect to hospital and operation effects. One formulation for such a model is

$$E\{F_{hi}\} = \sum_{k=1}^{t} \beta_k x_{hik} \quad \text{where} \quad \{x_{hi} = 1 \text{ for all } h, i\} ; \quad \left\{ \begin{array}{ll} x_{hi2} = 1 & \text{if } h = 2 \\ \quad\;\; 0 & \text{if } h = 1, 3, 4 \end{array} \right. \quad \left\{ \begin{array}{ll} x_{hi3} = 1 & \text{if } h = 3 \\ \quad\;\; 0 & \text{if } h = 1, 2, 4 \end{array} \right. \quad \left\{ \begin{array}{ll} x_{hi4} = 1 & \text{if } h = 4 \\ \quad\;\; 0 & \text{if } h = 1, 2, 3 \end{array} \right. ; \quad \left\{ \begin{array}{ll} x_{hi5} = 1 & \text{if } i = 1 \\ \quad\;\; 0 & \text{if } i = 0, 2, 3 \end{array} \right. \quad \left\{ \begin{array}{ll} x_{hi6} = 1 & \text{if } i = 2 \\ \quad\;\; 0 & \text{if } i = 0, 1, 3 \end{array} \right. \quad \left\{ \begin{array}{ll} x_{hi7} = 1 & \text{if } i = 3 \\ \quad\;\; 0 & \text{if } i = 0, 1, 2 \end{array} \right.$$

in which case $\beta_1$ represents a predicted value for operation 0 in hospital 1; $\beta_2$, $\beta_3$, and $\beta_4$ represent incremental effects for hospitals 2, 3, 4 respectively; and $\beta_5$, $\beta_6$, and $\beta_7$ represent incremental effects for operations 1, 2, and 3 respectively. The appropriateness of this model is confirmed by the non-significance of its goodness of fit statistic $Q(D.F.=9) = 6.33$ (which here corresponds to the hospital x operation interaction). Thus, certain hypotheses with respect to the parameters of this model can be tested in order to identify whether or not further model simplication can be undertaken. In this regard, the following hypotheses are of interest:

| Source of Variation | Hypothesis Formulation | D.F. | $Q_C$ |
|---|---|---|---|
| Hospitals | $\beta_2 = \beta_3 = \beta_4 = 0$ | 3 | 2.33 |
| Treatments | $\beta_5 = \beta_6 = \beta_7 = 0$ | 3 | 8.90 |
| Equality of Treatment Increments | $(\beta_6 - 2\beta_5) = (\beta_7 - 3\beta_5) = 0$ | 2 | 0.30 |
| Hospitals and Equality of Treatment Increments | $\beta_2 = \beta_3 = \beta_4 = 0$<br>$(\beta_6 - 2\beta_5) = (\beta_7 - 3\beta_5) = 0$ | 5 | 2.61 |

On the basis of these results, hospital effects can be removed from the model and treatment effects can be simplified to a single equal increment (linear) parameter. The specific structure of this model is shown in Table 2 together with corresponding predicted values and their standard errors. These predicted values indicate that the dumping syndrome severity functions $\{F_{hi}\}$ increase from the value of 0.20 for operation 0 to the value of 0.33 for operation 3 in increments of 0.04 per quarter of stomach removed (for each of the hospitals). Otherwise, the goodness of fit statistic for this model Q(D.F.=14) = 8.94 is non-significant ($\alpha$=0.25) and the test statistic for the equal increment parameter Q(D.F.=1) = 8.98 is significant ($\alpha$=0.01).

In summary, the structure of the model in Table 2 indicates that hospital effects can be ignored for the set of four hospitals which participated in this investigation during 1966-1968. Thus, it is plausible to extend this conclusion to all hospitals and all years and thereby argue that the equal increment relationship between the dumping syndrome severity functions $\{F_{hi}\}$ and extent of operation which was found to exist for these data could be generalized to this super-population.

i. Conclusions

i. Local inferences. There does exist a significant relationship between dumping syndrome severity and extent of operation after adjustment for (the partition of) hospital in the actual study population as defined in terms of the research design protocol, the four participating hospitals, and the 1966-1968 time period.

ii. Extended inferences. Since the study population in this application is relatively narrow in its definition, it is of interest to argue that its local inferences can be extrapolated to some larger target super-population of hospitals for which the same conclusions would be anticipated in future (or other) time periods. This point of view is supported by the non-significance of hospital effects and hospital x operation interaction. Otherwise, the extended inference analysis would have needed to take into account certain patient demographic and diagnostic covariables in order to produce a more complex super-population framework with respect to which hospital effects could be potentially ignored. In other words, if there is variation among hospitals, then it is not realistic to generalize local inferences for the self-selected (by their willingness and/or ability to participate) hospitals in this study to some larger population. However, if such variation can be statistically explained in terms of variation in patient populations with respect to certain covariables, then extended inferences are plausible for the stratified super-population corresponding to the multi-way cross-classification of these covariables and operation. Thus, from this type of point of view, the structure of the fitted model $\underset{\sim}{X}$ in Table 2 and its corresponding predicted values for the data from this investigation are considered to be of general descriptive interest with respect to the relationship between dumping syndrome severity and extent of operation. Otherwise, conclusions which are based on such results are subject to the same type of caution indicated for Example 1 in the sense of being inherently tentative until they receive further support by similarly designed studies or observed experience at other hospitals during future time periods. On the other hand, the practical importance of such qualifying statements is potentially reduced considerably for such experimental situations when they are conducted in a carefully controlled manner with strict adherence to the design protocol and strict maintenance of data quality control and when they include participating hospitals (or clinics) which reflect coverage of a broad range of patients in terms of geographic area, demographic characteristics, and diagnostic characteristics.

j. Other sources of examples for experimental data include

i. Experiments involving animals from certain types of breeding colonies;

ii. Experiments involving agricultural plots in certain judgmentally (as opposed to randomly) selected geographic areas;

iii. Experiments involving persons who are linked to certain institutions (schools, hospitals, criminal justice system) at certain judgmentally (as opposed to randomly) selected locations.

k. Summary statement concerning methodological issues. The most critical consideration underlying the interpretation of the analysis of experimental data is data quality as reflected by the extent to which there was strict adherence to the research design protocol. In this regard, potential sources of difficulty include protocol violations, missing data, and certain sources of measurement error (and/or bias). More specifically, if these types of data quality problems can be avoided (or managed in a substantively acceptable manner), then local inferences (concerning the randomized variable; e.g., operation type) can be undertaken in an assumption-free manner via the probabilistic structure (e.g., Model 0) induced on the data by the randomization component of the basic experimental design; and those concerning other variables can be undertaken in the hypothetical sense which was described with respect to Example 1. If there is interest in extending the scope of the conclusions of the local inference results, technical aspects of data analysis like variable scaling, variable selection, and model formulation (as discussed previously for this example and also in Koch, Freeman, and Lehnen [1976] and Higgins and Koch [1977]) become important so that the linkage between the sampled population and the target super-population is operationally defined in a sufficiently relevant manner.

## Appendix 1: Model 0

Let h = 1, 2, ..., q index a set of (s×r) contingency tables. Let i = 1, 2, ..., s index a set of sub-populations which are to be compared with respect to a particular response variable for which the outcome categories are indexed by j = 1, 2, ..., r. Let $n_{hij}$ denote the number of subjects (or study units) in the sample corresponding to the h-th table who are jointly classified as belonging to the i-th sub-population and the j-th response category. These frequency data can be summarized as shown in Table A1.

Table A1

| Sub-population | Response Variable Categories 1 | 2 | ... | r | Total |
|---|---|---|---|---|---|
| 1 | $n_{h11}$ | $n_{h12}$ | ... | $n_{h1r}$ | $N_{h1.}$ |
| 2 | $n_{h21}$ | $n_{h22}$ | ... | $n_{h2r}$ | $N_{h2.}$ |
| . | . | . | ... | . | . |
| . | . | . | ... | . | . |
| . | . | . | ... | . | . |
| s | $n_{hs1}$ | $n_{hs2}$ | ... | $n_{hsr}$ | $N_{hs.}$ |
| Total | $N_{h.1}$ | $N_{h.2}$ | ... | $N_{h.r}$ | $N_{h..}$ |

In this framework, $N_{hi.} = \sum_{j=1}^{r} n_{hij}$ denotes the marginal total number of subjects in the sample corresponding to the h-th table who are classified as belonging to the i-th sub-population, $N_{h.j} = \sum_{i=1}^{s} n_{hij}$ denotes the marginal total number of subjects in the sample corresponding to the h-th table who are classified as belonging to the j-th response category, and $N_{h..} = \sum_{i=1}^{s}\sum_{j=1}^{r} n_{hij}$ denotes the overall marginal total number of subjects in the sample corresponding to the h-th table. All of these quantities are assumed to be fixed constants rather than random variables. The types of situations where this type of assumption applies are

a. Observational and/or historical data from restricted populations as obtained in retrospective studies, case-control studies, etc.
b. Experimental design data from a strict randomization model point of view;
c. Product multinomial model sample data as described in Appendix 2 from a conditional distribution point of view.

The basic hypothesis of interest for this situation is

$H_0$: For each of the tables h = 1, 2, ..., q the response variable is distributed at random with respect to the sub-populations; i.e., the data in the respective rows of the h-th table can be regarded as a successive set of simple random samples of sizes $\{N_{hi.}\}$ from a fixed population corresponding to the marginal total distribution of the response variable $\{N_{h.j}\}$.

Under the hypothesis $H_0$, the following probability model characterizes the distribution of the $\{n_{hij}\}$.

$$\Pr(\{n_{hij}\}|H_0) = \prod_{h=1}^{q} \frac{\prod_{i=1}^{s} N_{hi.}! \prod_{j=1}^{r} N_{h.j}!}{N_{h..}! \prod_{i=1}^{s}\prod_{j=1}^{r} n_{hij}!}$$

From the structure of this model, it follows that

$$m_{hij} \equiv E\{n_{hij}|H_0\} = N_{hi.} N_{h.j}/N_{h..}$$

$$v_{h,ij,i'j'} = \mathrm{Cov}\{n_{hij}, n_{hi'j'}|H_0\} =$$

$$\frac{N_{hi.} N_{h.j} (\delta_{ii'} N_{h..} - N_{hi'.})(\delta_{jj'} N_{h..} - N_{h.j'})}{N_{h..}^2 (N_{h..} - 1)}$$

where $\delta_{ii'} = \begin{cases} 1 \text{ if } i = i' \\ 0 \text{ if } i \neq i' \end{cases}$ and

$\delta_{jj'}$ is similarly defined.

Let $\underset{\sim}{n}_h$ denote the vector of observed frequencies $\{n_{hij}\}$. Let $\underset{\sim}{m}_h$ denote the vector of hypothesis based expected frequencies $\{m_{hij}\}$, and let $\underset{\sim}{V}_h$ denote the hypothesis based covariance matrix $\{v_{h,ij,i'j'}\}$. Let $\underset{\sim}{A}$ be an [(r-1)(s-1) x rs] matrix which is rank independent of within table response sum and sub-population sum vectors (e.g., $\underset{\sim}{A}$ is the Kronecker product of any response contrast basis with any sub-population basis). Then, it follows that an appropriate test statistic for $H_0$ in a total sense is

$$Q_T = \sum_{h=1}^{q} \underset{\sim}{d}_h' \underset{\sim}{A}' \{\underset{\sim}{A} \underset{\sim}{V}_h \underset{\sim}{A}'\}^{-1} \underset{\sim}{A} \underset{\sim}{d}_h$$

$$= \sum_{h=1}^{q}\sum_{i=1}^{s}\sum_{j=1}^{r} \left(\frac{N_{h..} - 1}{N_{h..}}\right)\left\{\frac{(n_{hij} - m_{hij})^2}{m_{hij}}\right\}$$

$$= \sum_{h=1}^{q} \left(\frac{N_{h..} - 1}{N_{h..}}\right) Q_{P,h}$$

where $\underset{\sim}{d}_h = (\underset{\sim}{n}_h - \underset{\sim}{m}_h)$ and $Q_{P,h}$ is the Pearson Chi-Square statistic for the h-th table. Under $H_0$, $Q_{P,h}$ asymptotically has the chi-square distribution with D.F. = (r-1)(s-1). Thus, if all the $\{N_{h..}\}$ are sufficiently large, both $Q_T$ and $Q_{TP} = \sum_{h=1}^{q} Q_{P,h}$ have approximate chi-square distributions with D.F. = q(r-1)(s-1).

On the other hand, if many of the $\{N_{h..}\}$ are small even though the overall sample size $N = \sum_{h=1}^{q} N_{h..}$ is large, then $Q_T$ (and $Q_{TP}$) are no longer appropriate for testing $H_0$. In this situation, the Cochran-Mantel-Haenszel type of statistic can be used. These have the form

$$Q_{CMH} = \underset{\sim}{d}_{\cdot} \, \underset{\sim}{V}_{\cdot}^{-1} \, \underset{\sim}{d}_{\cdot}$$

where $\underset{\sim}{d}_{\cdot} = \sum_{h=1}^{q} \underset{\sim}{A}_h \underset{\sim}{d}_h$ and $\underset{\sim}{V}_{\cdot} = \sum_{h=1}^{q} \underset{\sim}{A}_h \underset{\sim}{V}_h \underset{\sim}{A}_h'$. Under $H_0$, $Q_{CMH}$ has approximately the chi-square distribution with D.F. = (r-1)(s-1). Otherwise, it can be noted that $Q_{CMH}$ is directed at average partial association alternatives in the sense that if certain elements of $d_h$ are consistently positive (or negative) across the tables h = 1, 2, ..., q, then these quantities reinforce one another when combined to form $d_{\cdot}$. Also, the fact that significance of $Q_{CMH}$ is evaluated relative to D.F. = (r-1)(s-1) rather than q(r-1)(s-1) represents another aspect of this approach that potentially permits gains in statistical power here.

In some applications, the response categories may be ordinally scaled, in which case location shifts with respect to this scaling often represent the primary types of alternatives of interest. Thus, it becomes advantageous to target the statistics $Q_T$ and $Q_{CMH}$ on certain mean score functions of the type $F_{hi} = \sum_{j=1}^{r} a_{hj} n_{hij}$ where the $\{a_{hj}\}$ represent a reasonable set of numerical values which have been assigned to the set of ordinally scaled response categories. For this purpose, the basic formulas given for $Q_T$ and $Q_{CMH}$ remain essentially the same except that the matrix $\underset{\sim}{A}$ is allowed to vary across tables in the form $\underset{\sim}{A}_h$ and each $\underset{\sim}{A}_h$ is an [(s-1) × rs] basis of sub-population contrast space with respect to the specific linear combination of response categories that pertain to the functions $\{F_{hi}\}$ within that table. In view of the reduced dimension of $\underset{\sim}{A}$ which these modifications involve, $Q_T$ has asymptotically the chi-square distribution with D.F.=q(s-1) and $Q_{CMH}$ has asymptotically the chi-square distribution with D.F.=(s-1).

Finally, if both the response categories and the sub-population categories are ordinally scaled, then certain types of correlation alternatives are often of primary interest. In these situations, it is advantageous to target $Q_T$ and $Q_{CMH}$ on a single function of the type $F_h = \sum_{i=1}^{s} \sum_{j=1}^{r} c_{hi} a_{hj} n_{hij}$ for each table where the $\{c_{hi}\}$ represent a reasonable set of numerical values which have been assigned to the ordinally scaled sub-population categories. Otherwise, the formulas for $Q_T$ and $Q_{CMH}$ remain essentially the same as originally given, except that $\underset{\sim}{A}$ is allowed to vary across tables in the form $\underset{\sim}{A}_h$, each $\underset{\sim}{A}_h$ has only a single row whose elements are the respective products $\{c_{hi} a_{hj}\}$, and the asymptotic chi-square distributions for $Q_T$ and $Q_{CMH}$ have D.F.=q and D.F.=1 respectively.

For further discussion of Model 0 and the various types of statistics which are of interest with respect to it, see Landis et al. [1977].

## Appendix 2: Model 0

For the same general framework described in Appendix 1, the $\{n_{hij}\}$ are assumed to follow the product multinomial distribution.

$$\Pr(\{n_{hij}\}) = \prod_{h=1}^{q} \prod_{i=1}^{s} \prod_{j=1}^{r} \frac{N_{hi.}! \, \pi_{hij}^{n_{hij}}}{n_{hij}!}$$

where $\pi_{hij}$ represents the probability that a randomly selected subject from the (hi)-th sub-population is classified in the j-th response category. The type of situations where this type of assumption is appropriate are

i. Stratified simple random sampling from an infinite super-population where the strata correspond to the qs cells of the h vs i cross-classification;
ii. Simple random sampling from an infinite super-population from a conditional distribution point of view, in which case h vs i is a domain cross-classification;
iii. Certain mixtures of (i) and (ii) where h accounts for the variables in terms of which the stratification cross-classification is defined and i accounts for the variables in terms of which the domain cross-classification is defined.

Let $p_{hij} = (n_{hij}/N_{hi.})$ denote the proportion of subjects in the sample from the (hi)-th sub-population that are classified in the j-th response category. The $\{p_{hij}\}$ represent unrestricted maximum likelihood estimates of the $\{\pi_{hij}\}$. Let $\underset{\sim}{p}$ denote the vector of $\{p_{hij}\}$ and let $\underset{\sim}{\pi}$ denote the vector of $\{\pi_{hij}\}$.

Depending on the nature of the situation under consideration, certain aspects of the response distribution within each sub-population and/or its relationship to the sub-populations can be formulated in terms of functional transformations $\underset{\sim}{F}(\underset{\sim}{\pi})$, and the extent to which there is variation among such functions can be characterized by linear regression models of the type $\underset{\sim}{F}(\underset{\sim}{\pi}) = \underset{\sim}{X}\underset{\sim}{\beta}$. Thus, the principal objectives of statistical analysis include the estimation of the model parameters $\underset{\sim}{\beta}$ and the corresponding predicted values they imply for $\underset{\sim}{F}(\underset{\sim}{\pi})$, statistical tests for hypotheses involving $\underset{\sim}{\beta}$, and statistical tests for the goodness of fit of the model $\underset{\sim}{X}$. For this purpose, two general approaches which have wide applicability to many specific problems of this type are maximum likelihood methods as discussed in Bishop, Fienberg, and Holland [1975] and weighted least squares asymptotic regression as discussed in Grizzle, Starmer, and Koch [1969], and Koch et al. [1977].

## Appendix 3: Model 2

For many types of research investigations, data are obtained via probability random samples with complex designs. Some strategies for their analysis relative to the sampled population are discussed in Koch et al. [1975]. However, super-population issues are philosophically more difficult here because the nature of the hypothetical selection process is not necessarily well-defined. For this reason, one simplistic approach is to adopt a Model 0 or Model 1 point of view for this situation.