

## **Collecting and Analyzing Data**

### ***Data Tabulation***

The following chapter is excerpted from *Designing HIV/AIDS Intervention Studies: An Operations Research Handbook*, Andrew Fisher and James Foreit, 2002, Washington, DC: Population Council. ([More on OR Handbook](#))

# TABULATION OF DATA

In your proposal, you should discuss editing and tabulating data immediately after data collection procedures. Although qualitative methods are being increasingly used in operations research, most OR studies still involve quantitative analysis that requires statistical manipulation of the information collected.

First, you need to convert the information into a form that will allow it to be analyzed. Second, you must specify the statistical manipulations to be performed. Finally, you need to present the important findings resulting from these manipulations in a report or series of reports.

## Preparing Tabulations

Any recently produced desktop computer probably has the hardware capability needed to process an operations research data set. However, unless the computer you use is located at a research organization or a health program evaluation unit, it may not have the software needed for statistical analysis. This is not a problem when data consist only of service statistics from a small number of service delivery points, or when modeling or conducting a cost analysis; in both cases, spreadsheets are adequate for analyzing OR data. However, it's more likely that you will need to perform many statistical tests, analyze survey data, or work with a very large data set and will need more powerful statistical software.

*Epi Info* is a statistical package that is available free of charge from the U.S. Centers for Disease Control and Prevention (CDC). It has features for processing and analyzing data, including survey data. Although it is a basic package, it has all the features necessary for analyzing most OR studies. More powerful (and expensive) software packages include *SPSS*, *STATA*, and *SAS*, all of which require training to use. In deciding on software, it is wise to select a program that is widely used in your country or in your organization, since it will then be much easier to find technical support and consultants.

## Data Coding

All statistical packages include data entry features. But before you begin to enter data, you must transform the raw information for tabulation and analysis. Nonnumerical data that are to be analyzed quantitatively must be converted into numerical codes. If your data gathering instrument uses mainly closed questions (a question with a limited number of possible predetermined responses, such as “yes” or “no”), the best approach is to **precode** the instrument. Thus, the question would appear with the numeric codes for the responses already printed on the instrument, as shown in the example below:

QUESTION No.	QUESTION	RESPONSE	SKIP TO
110	Where did you get your HIV test?	Hope Hospital =1	
		Central Hospital =2	Q115
		Town Clinic =3	Q116
		Military Camp =4	Q117
		Other (Specify) =9	
	Nonresponse		

If, in response to question 110, the respondent states that he received his last HIV test at Central Hospital, the interviewer would circle the number 2. If the answer is the military camp, the interviewer would circle 4. Before beginning computerized data entry, you should check all questionnaires to make sure that the interviewers have recorded a response to each question.

If the number of categories for a particular variable (including, if relevant, codes for “nonresponse,” “not applicable,” “don’t know,” and “other”) is less than 10, numerical codes should be single-digit numerals. If the total possible number of categories is between 10 and 99, two-digit codes should be used instead. For some variables, it may be necessary to use three-digit, four-digit, or even larger codes; for example, calendar dates typically require four or more digits.

## Data Entry and Editing

Coded data need to be entered into the computer with a minimum of typing errors and then edited to correct any errors in the data. In entering data, the researcher should use the data verification procedures available with most statistical packages. In verification, the same data are entered twice. The verification program indicates discrepancies in the numbers entered. In the example above, the first data entry clerk might have entered the number 3. However, the second time the response is entered, it may be entered as 1. When such discrepancies occur, the program signals the data entry person to check the data entry form for the correct number.

In addition to verification, the researcher should check for the following types of errors:

- **“Illegal” codes:** Values that are not specified in the coding instructions. For example, a code of “7” in question 110 above would be an illegal code. The best way to check for illegal codes is to have the computer produce a frequency distribution and check it for illegal codes.

- **Omissions:** For example, a failure by an interviewer to follow correctly the SKIP instructions in a questionnaire. This would be the case in question 110, if the interviewer failed to skip to question 115 after a response of “Central Hospital.”
- **Logical inconsistencies:** For example, a respondent whose current age is less than her age at marriage.
- **Improbabilities:** For example, a 25-year-old woman with ten living children.

Once you find errors, check the original data forms to make the necessary corrections. Most coded data can be edited on the computer, but **field editing** should always be done by supervisors whenever there is a chance that the error can be corrected by talking with the data gatherer or perhaps re-interviewing the respondent for clarification.

## Variable Transformations

Once data have been entered into the database, it is often necessary to transform variables. The transformations may constitute the entire analysis of the study, but far more often data transformations are done to permit subsequent analyses.

For instance, instead of having the questionnaire record the respondent’s age, the questionnaire may record only the month and year of birth. If age is a variable to be studied, it can be obtained simply by having the computer subtract the month and year of birth from the month and year of the interview.

This transformed variable might be transformed even further for certain kinds of additional analysis. For example, if you want to cross-tabulate age by other variables, it is preferable to limit the age distribution to relatively few age categories (usually five- or ten-year categories) or even to dichotomize (for example, ages 15–29 and ages 30 or more). You can use several methods to transform variables, the most common of which are listed below.

### RECODES

In recoding, category labels are changed. This technique is used to “collapse” large numbers of variable categories into smaller numbers. For example, single years of age can be collapsed and transformed into age categories, such as ages 15–19, 20–24, and 25–29.

### COUNTS

If you are collecting information on whether the respondents have ever used any of eight services for persons with HIV/AIDS, you might want to count the number of services ever used by each respondent. Thus, you could generate a new variable that might be called “Number of Services Ever Used.”

### CONDITIONAL TRANSFORMATIONS

When the nature of the transformation of one variable depends on the second variable, conditional transformations may be useful. For instance, suppose you asked respondents three questions:

- Did you hear the partner reduction radio message in July?
- How many casual sex partners did you have between April and June?
- How many casual sex partners did you have between August and October?

With the information from these three questions, you could then create a new variable called “Partner Reduction among Persons Exposed to Radio Message.” This can be done by subtracting the number of partners in question 3 from the number in question 2. But you would do this only if the answer to question one is “yes.”

### OTHER MATHEMATICAL TRANSFORMATIONS

Calculating age from the date of birth and the date of the interview is an example of a mathematical transformation. Another example is obtaining an HIV prevalence rate by dividing the number of HIV-positive individuals in a community by the number of residents.

## What To Do: Coding and Editing

1. Be sure to check the availability of computers, statistical packages, and programming assistance before you start.
2. Indicate in the proposal the checking and editing that you will do.
3. Precode your questionnaire or other instrument.
4. Prepare a codebook that labels and specifies the meaning of each numerical value of all variables in your database.
5. Plan for editing during field work.
6. Verify the accuracy of the numerical values entered into the database.
7. When data entry has been completed, check for illegal codes, omissions, inconsistencies, and improbabilities before analyzing your data.
8. Begin to perform basic variable transformations.