WHAT IT IS

Return to Table of Contents

Frequency distributions summarize and compress data by grouping it into classes and recording how many data points fall into each class. That is, they show how many observations on a given variable have a particular attribute. For example, a survey is taken of 50 people's favorite color. The frequency distribution might indicate 15 people selected green, 12 blue, 6 red, 7 yellow, and 10 purple. Converting these raw numbers into percentages would then provide an even more useful description of the data.

The frequency distribution is the foundation of descriptive statistics. It is a prerequisite for both the various graphs used to display data and the basic statistics used to describe a data set -- mean, median, mode, variance, standard deviation, and so forth. Note that frequency distributions are generally used to describe both nominal and interval data, though they can describe ordinal data.

WHEN TO USE IT

A frequency distribution should be constructed for virtually all data sets. They are especially useful whenever a broad, easily understood description of data concentration and spread is needed. Most data provided by third parties are grouped into a frequency distribution.

HOW TO PREPARE IT

Regardless of whether manual or automated methods are used to prepare a frequency distribution, it is usually necessary to code data numerically to facilitate further data analysis. This makes creating a data dictionary which defines the numeric codes used to identify data categories necessary. For example, assume that an auditor/evaluator wants to classify both demographic data and information on the opinion of entity staff on a particular policy. A data dictionary for use with computer software might resemble the following:

Variable Name	Code	Field Width	Field Type
Division	Actual Division	20	Alphanumeric
Age	Age in Years	3	Numeric
Gender	1 = Male 2 = Female	1	Numeric
Salary Range	1 = $ 0 - 20,000 2 = $20 - 30,000 3 = $30 - 50,000 4 = Over 50,000	5	Numeric
Policy Opinion	1 = Excellent 2 = Good 3 = Fair 4 = Poor	1	Numeric

It is also necessary to determine how many classes one should use for the frequency distribution. Selecting a number of classes is not as arbitrary as may first appear. If data are nominal, simply list all possible classes (i.e. categories) into which a data point might fall. If data are interval, the table below can function as a rule of thumb:

Number of Observations	Number of Classes
Under 50	5 - 7
50 - 200	7 - 9
200 - 500	9 - 10
500 - 1,000	10 - 11
1,000 - 5,000	11 - 13
5,000 - 50,000	13 - 17
Over 50,000	17 - 20

If the data are nominal, a contingency table may be useful. Contingency tables are discussed near the end of this module.

Preparation

The steps in preparing frequency distributions manually are as follows:

- Collect raw data from entity records, interviews, surveys, etc.
- If data are nominal, list the classes into which a data point might fall. If data are interval, select an appropriate number of data classes.
- Calculate the absolute frequency of each class, i.e. the raw number of data points in each class. Note that the sum of all absolute frequencies must equal the sample size.
- Calculate the relative frequency by dividing the absolute frequency by the sample size. This reveals the proportion or percent of data points in each class. Note that the sum of all relative frequencies must be 1.
- Calculate the cumulative frequency for each class by adding the number or proportion/percentage of data points in that class to similar quantities for all preceding classes. If you accumulate the number of data points, the last number should equal the sample size. If you accumulate the proportions/percentages, the last number should be 1 or 100 percent.

While it is possible to prepare a frequency distribution by hand, it is preferable to use software such as Quattro Pro, Lotus, SAS, or a statistics package.

HOW TO REPORT IT

The results of frequency distributions and contingency tables can often be used in reports. Percentages are usually reported rather than raw counts to give readers more context.

Using the variables listed in the data dictionary example on page 1 of this module, a brief write-up might be:

> "The survey was answered by 55 employees, of whom 47 percent were male and 53 percent female. Minorities comprised 18 percent of survey respondents. Only 43 percent of employees responding to the survey generally agreed with the policy. Responses varied along demographic lines. Women were more likely than men to view the policy favorably (45 percent to 38 percent). Younger workers, those under age 35, were also more likely to view the policy as beneficial than were older workers (47 percent to 41 percent)."

A NOTE ON CONTINGENCY TABLES

Contingency tables provide more information about a data set than do simple frequency distributions, particularly when data are nominal. In the case of the survey of staff opinions on a given policy noted on page 1 of this module, those evaluating the survey might want to compare the responses of women to those of men. Contingency tables allow frequency counts to be broken up, examined in multiple dimensions, and then tested for independence and statistical significance via the χ^2 statistic. The following is an example of a 2-by-2 contingency table which examines the opinions male and female employees on the aforementioned policy.

Opinion	**Male**	**Female**	**Total**
Unfavorable	10 (38%)	13 (45%)	23 (42%)
Favorable	16 (62%)	16 (55%)	32 (58%)

ADVANTAGES

Frequency distributions can:

- condense and summarize large amounts of data in a useful format
- describe all variable types
- facilitate graphic presentation of data
- begin to identify population characteristics
- permit cautious comparison of data sets

DISADVANTAGES

Frequency distributions can:

- reveal little about the actual distribution, skew, and kurtosis of data
- be easily manipulated to yield misleading results
- de-emphasize ranges and extreme values, particularly when open classes are used (e.g., “over 65,” “under $15,000" etc.)