1st BJH

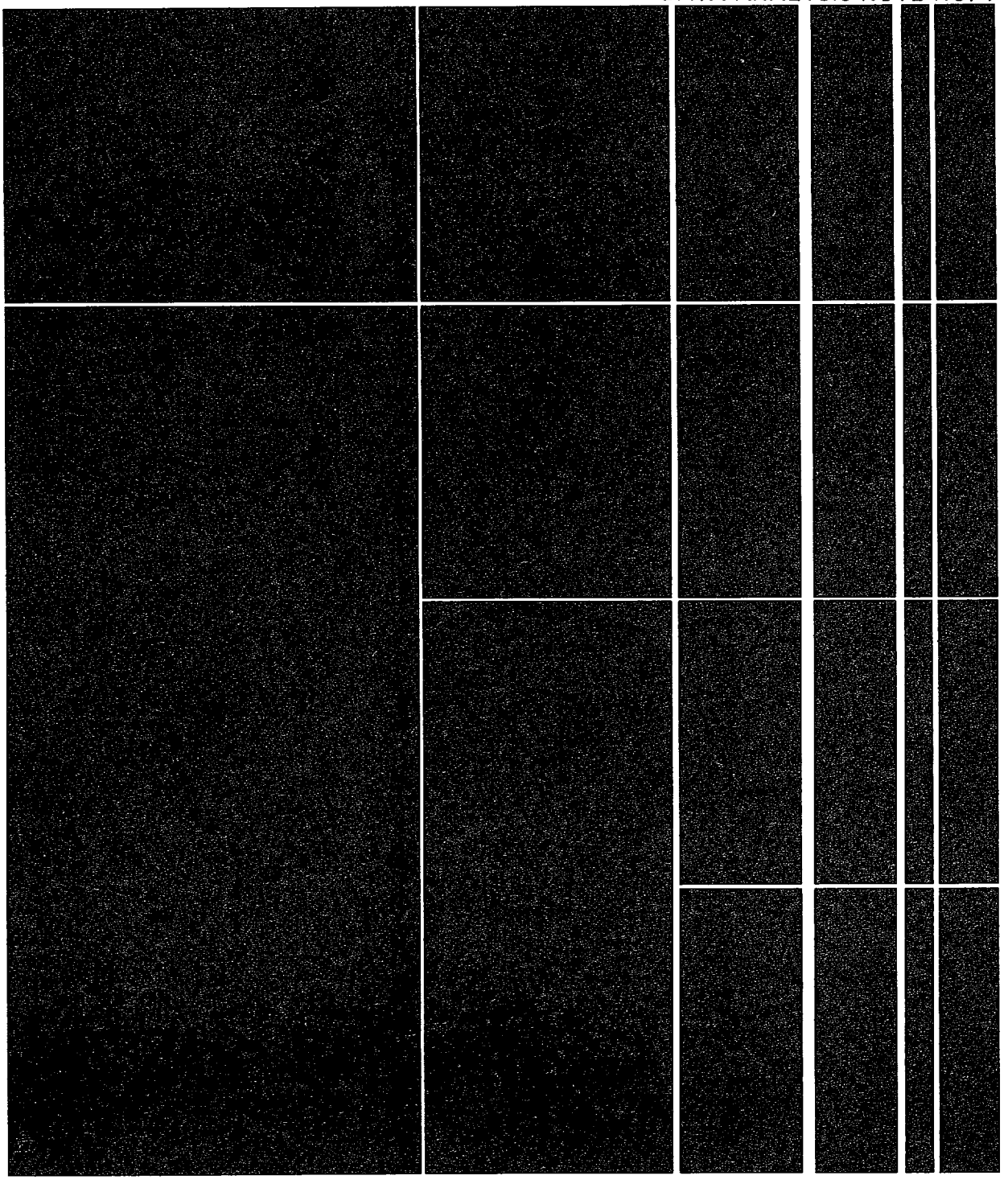
8938

EXPLANATIONS OF SOME SPSS PROCEDURES

GAO/150

FEDERAL PERSONNEL AND COMPENSATION DIVISION
SYSTEMS ANALYSIS GROUP

DATA ANALYSIS NOTE NO. 1



003330

Other

FOREWORD

As part of its consulting function in FPCD, the Systems Analysis Group is publishing a series of booklets designed to enhance the auditor's ability to make maximum use of data collected during an assignment. These booklets are intended for use by people with little or no statistical or computer background. They will describe the availability, use, and interpretation of statistics, computer programs, and scientific procedures which can be used to increase effectiveness and efficiency of FPCD assignments.

This booklet and its appendix are the first in the series. Together they are designed to facilitate the auditor's task of interpreting computer output. The booklet itself describes some of the statistics, tables, and graphs available in the most widely used package of social science computer programs. The appendix contains actual examples of output from these programs and should be referred to as noted in the booklet.

The idea for this particular booklet grew from the Systems Analysis Group's experience with interpreting computer output for others. The group felt that a simple reference manual, such as this one, would reinforce verbal explanations about the meaning and significance of information contained in printouts. Perhaps, more importantly, the booklet may help the auditor to see different ways of analyzing data and displaying the results of that analysis.

Please feel free to contact the Systems Analysis Group for additional information. This booklet and its appendix were prepared by Nancy Simmons.

H. L. Krieger
Director, Federal Personnel
and Compensation Division

Table of Contents

Introduction

The descriptions presented in this booklet involve types of analyses that are provided by a computer package called the Statistical Package for the Social Sciences (SPSS). This statistical package is designed to analyze social science data and is one of the most widely used statistical packages. The main requirement for using SPSS is the existence of units of information referred to as data. Information, or data, for analysis on SPSS can be obtained in many forms from many sources. Questionnaires, structured interviews, and pro forma schedules are all tools for gathering information. Accessing computerized inventory, personnel, and other types of records from GAO or other agencies are a few more ways of gathering data.

One reason SPSS is so widely used is that it permits great flexibility in handling and displaying data. First, both alphabetic information, such as names and addresses, and numeric information, such as social security numbers and percentages, can be used with SPSS. Second, with SPSS, units of information can be combined or modified, new information can be created from the original information, and the information can be divided into subsets for analysis. Finally, the programmer has a wide choice of tables and graphs to use for displaying the data. Other packages are available, but the Systems Analysis Group can advise which package is best to use for the anticipated analyses.

Each chapter in this booklet describes a different program available with SPSS. There are more programs available, but these are some of the most commonly used ones. The programs called CONDESCRIPTIVE, BREAKDOWN, and CROSSBREAK give general statistics on a set of data such as the average value, the highest value, the lowest value, etc. The frequency of occurrence of different answers or values for one unit of information is given by the program FREQUENCIES. The remaining programs, CROSSTABS, T-TEST, and PEARSON CORR, give statistics which show relationships between two or more units of information.

On each page of computer output, the first few lines contain the same information. (Refer to page 1 of the appendix for examples.) The first line contains a heading, such as SOME SPSS OUTPUT (on page 1 of the appendix), which is designated by the programmer. On the same line and to the right are the date the output was generated and the page number. The heading and date identify the printout and distinguish it from other printouts. On the second line, the coded name of the information file created is followed by

that file's creation date. Sometimes the data is divided into subfiles or subdivisions so that they may be analyzed separately, such as dividing GAO personnel information into subfiles by divisions or offices. In these instances, the subfile list will be printed beneath the name of the information file (see page 6 of the appendix).

The next five chapters will describe the remainder of the pages of output which differ depending on the procedure and statistics requested. Throughout these chapters the following terminology and conventions are used:

--The term case refers to a complete set of information. It is also called a record. For example, answers to one questionnaire form a case, or information from one personnel file forms a case.

--The term variable refers to a single item of information. A variable has many values, but only one value per case. For example, the variable age can have values of zero or higher but there will be only one age for each case.

--The term data refers to the variables and their values as well as to the values on the output.

--Run, value, and variable labels are usually printed but may be omitted from one or more sets of output if desired.

--Variable names are used instead of actual names for data because of restrictions in formation and length of the names as well as convenience to the programmer.

--MISSING values or observations are the number of cases which were omitted from the calculations. These are omitted because values for the variable or variables analyzed were invalid, left blank, illegible, not applicable, etc.

--Since data used is usually from a sample, or representative section of a larger group, the output from SPSS should be used as an indication of how the whole group looks. It cannot be used to definitely conclude anything about the whole group; only conclusions about that sample are valid.

This booklet describes six of the SPSS programs. There are other programs available which the Systems Analysis Group can suggest when applicable. Anyone interested in learning more details about SPSS should feel free to contact

members of the Systems Analysis Group in FPCD or look over a copy of the SPSS manual. Version 6.0 of the manual can be found in the GAO library under SPSS by Norman Nie (HA33.N48).

CONDESCRIPTIVE

This SPSS program is designed to describe the degree of dispersion or spread of answers or values for a unit of information. The answers are collected by the program from the original information, and statistics are computed regarding the range of the answers or how close the answers are to one another in value. This program is most effective when used with information for which an average value makes sense. For example, averages for age, cost, or years of service make sense; but averages for sex or social security number do not.

An example of the output resulting from this program is found on page 1 of the appendix. Statistics are given for each of five units of information; and each set of statistics is divided by dashed lines (see page 1 of the appendix).

The first line of each section lists the coded name and a definition of the information. The next three lines list the names of the statistics and numbers or values associated with each statistic. Finally, the last line in each section gives information about the number of cases analyzed.

The information used to produce page 1 of the appendix was gathered by using a questionnaire. This questionnaire was distributed to 1,070 enlisted personnel assigned to five ships that were undergoing overhaul or conversion. Questionnaire responses indicated the satisfaction and utilization of the crew members. For this page of output, a subpopulation of 1,064 of the responses was selected and analyzed. The following table gives interpretations and examples of statistics listed in the first section on the output. Reference numbers refer to page 1 of the appendix.

NAME	VALUE	DESCRIPTION AND/OR INTERPRETATION	REFERENCES OR EXAMPLES
VARIABLE	Q1	PAY GRADE - the statistics in the first section refer to pay grade. Q1 is the coded name for pay grade.	See note (1) on page 1 of the appendix.
MEAN	3.992	The sum of all the responses to pay grade divided by the number of answers; i.e., the average answer. Note: this may not be a good measure of central tendency if the responses are not evenly distributed around the mean.	See note (2). The average pay grade is approximately E-4 (the respondents were all enlisted, so 4 means E-4.)
VARIANCE	2.440	A measure of how close in value the answers are to the mean answer. Standard deviation is usually a better measure of dispersion.	See note (3). Since the variance is a little smaller than the mean, the values of pay grade are not very dispersed.
STD DEV	1.562	Standard deviation is another measure of how closely the answers cluster around the mean. This statistic is in the same units of measure as the mean. STD DEV = $\sqrt{\text{VARIANCE}}$.	See note (8). The standard deviation is fairly small relative to the mean indicating little dispersion in values for pay grade.
STD ERROR	.048	Standard error indicates the accuracy of the value of the mean.	See note (5). Since standard error is small, the mean is fairly accurate.
MINIMUM	1.000	The lowest valued answer.	See note (7). E-1 was the lowest pay grade listed.

NAME	VALUE	DESCRIPTION AND/OR INTERPRETATION	REFERENCES OR EXAMPLES
MAXIMUM	9.000	The highest valued answer.	See note (10). E-9 was the highest pay grade listed.
RANGE	8.000	The difference between MAXIMUM and MINIMUM. A measure of dispersion of the answers.	See note (4). The highest and lowest answers are 8 units apart.
SKEWNESS	.480	A measure that describes the distribution of the answers and what kind of curve would represent them. For a more detailed description and interpretation consult the Systems Analysis Group.	See note (9). The values of pay grade are fairly evenly distributed about the mean.
KURTOSIS	-.265	This is another measure of the distribution of the answers and should be used along with SKEWNESS. For a more detailed description and interpretation consult the Systems Analysis Group.	See note (6).
VALID OBSERVATIONS	1,063	The number of cases analyzed.	See note (11).
MISSING OBSERVATIONS	1	The number of cases omitted from analysis.	See note (12). Cases with no answer for pay grade were designated MISSING.

FREQUENCIES

This SPSS program is designed to give frequency of occurrence statistics for a unit of information. The answers are collected and sorted by their values; then a table is generated containing a raw count and percentage statistics for each value. This program is most effective when used with information having a relatively small number of values, such as sex or month of birth. Items such as age and salary or information using decimals will usually have so many values that counts for each value will not be very useful.

Several examples of output resulting from this program are found in the appendix on pages 2 through 6. The information used to produce these five pages of output came from the questionnaire mentioned in the preceding section. All of the 1,070 cases were analyzed on the first two pages of output, while a subpopulation of 1,064 cases was selected and analyzed on the last three pages. The first line on each of these contains the coded name and a definition of the information (see note (1) on page 2). The last line of each page lists VALID CASES and MISSING CASES (see note (9) on page 2). MISSING CASES are those that contain values which should be left out of some calculations (see note (8)). VALID CASES are cases which are included in the analyses. The rest of the page varies depending on the format requested.

There are basically three formats for output from the FREQUENCIES program. The first is a tabular format shown on pages 2, 3, and 4 of the appendix. Page 2 of the appendix involves numerical data, page 3 involves alphabetical data, and page 4 demonstrates an optional format in which the output is printed on an 8-1/2 x 11 inch space to the left of the page. Otherwise, these three examples are basically the same. The following page interprets the table on page 2 of the appendix. The coded name and description of the information being analyzed are noted as (1) on page 2 of the appendix.

COLUMN HEADINGS	DESCRIPTIONS	EXAMPLES
CATEGORY LABEL	This column gives labels for the numerically coded values of the answers.	See note (2). The number 1 in the CODE column corresponds to a marital status answer of "MARRIED," 2 corresponds to "SINGLE," etc.
CODE	The coded answers or values to the information being analyzed.	See note (3). The answers for MARITAL STATUS have been coded as 0, 1, 2 and 3.
ABSOLUTE FREQ	This is the number of times a given value is encountered, or the number of respondents which listed each value.	See note (4). 433 respondents listed their MARITAL STATUS as MARRIED.
RELATIVE FREQ	This is the number in the ABSOLUTE FREQ column expressed as a percentage of total cases.	See note (5). The 433 married respondents represent 40.5% of the 1,070 cases.
ADJUSTED FREQ	This number is usually the same as RELATIVE FREQ. However, it is often convenient to delete some values before computing the percentages. These values are called MISSING. Any missing values appear at the end of the ADJUSTED FREQ column.	See note (6). The 433 married respondents represent 40.5% of the valid cases. Notice that NO ANSWER for marital status was declared MISSING (see note (8)).
CUM FREQ	This is the cumulative total of the percentages in the ADJUSTED FREQ column.	See note (7). 93.4% is the total of 40.5% and 52.9%.

The second type of format for FREQUENCIES output is shown on page 5 of the appendix. The CATEGORY LABEL for each CODE is omitted and the values of CODE are listed horizontally rather than vertically. The statistics given are:

CODE - the coded values to the data being analyzed (see note (1) on page 5).

FREQ - same as ABSOLUTE FREQ (see note (2) on page 5).

ADJ PCT - same as ADJUSTED FREQ (see note (3) on page 5).

CUM PCT - same as CUM FREQ (see note (4) on page 5).

Beneath the frequencies and percentages is a section which gives the coded value and frequency of MISSING DATA (see note (5) on page 5).

The third type of format for FREQUENCIES is shown on page 6 of the appendix and is called a histogram. It pictorially represents frequencies by graphing the counts for each value of the data item by using asterisks (see note (1) on page 6). The horizontal axis (see note (2) on page 6) is the frequency and the vertical axis (see note (3) on page 6) is the coded value. The MISSING values are denoted along the vertical axis (see note (4) on page 6). CATEGORY LABELS are listed under the line they label (see note (5)); and the number in parentheses to the right of each line is the same as ABSOLUTE FREQ (see note (6)).

In addition to frequency counts and percentages, some other statistics are available. These are listed after the table on page 4 of the appendix. All but two of these statistics were described in the chapter on CONDESCRIPTIVE. The two additional ones are:

STATISTIC	DESCRIPTION	EXAMPLE
MODE	The value that occurs most often.	See note (1) on page 4 of the appendix. The value 2 occurred more often than 0, 1, or 3.
MEDIAN	The value of the middle number when all the answers are listed in ascending order. There are as many answers greater than the median as there are less than the median.	See note (2) on page 4 of the appendix. The middle answer is between 1 and 2; it is 1.679.

Please refer to the CONDESCRIPTIVE chapter of this booklet for interpretations of the other statistics.

CROSSTABS

This SPSS program generates tables describing certain relationships between two or more units of information. The tables can be used to indicate tendencies such as whether married respondents to a questionnaire are more likely to own a house than single respondents, whether salary is higher for older respondents than for younger ones, etc.

Examples of the output resulting from this program are found on pages 7 through 10 of the appendix. The information used to produce this output also came from the questionnaire distributed to enlisted personnel on ships undergoing overhaul. All the 1,070 cases were chosen for analysis in the first table, a subpopulation of 1,031 cases was chosen for analysis in the next two tables combined, and a subpopulation of 1,064 cases was chosen for analysis in the last table. At the top of each page of output (see page 7 of the appendix) the coded name and a definition of the information being compared is given. The last line of the output, which follows the table, gives the NUMBER OF MISSING OBSERVATIONS (see note (1) on page 7). This is the number of cases that involved values which the programmer designated MISSING so that they would be omitted from the table. In the case of the table on page 7, the 2 cases were respondents who did not answer either for the information coded Q2 or for the information coded Q4 or both.

The following chart shows how to interpret the figures in the upper left-hand box as well as the row and column totals of the table on page 7 of the appendix. The notes given refer to this table.

LABEL	DESCRIPTION	EXAMPLE
Q2	MARITAL STATUS - the coded name for the information whose values will be listed down the side of the table.	See notes (2). The values for MARITAL STATUS (1, 2, 3) are listed down the side of the table.
Q4	HAVE HOMEPORT RESIDENCE? - the coded name for the information whose values will be used across the top of the table.	See notes (3). The values for whether or not the person has a homeport residence (1,2) are listed across the top of the table.
ROW TOTAL	The number of respondents in each row is listed to the right of that row.	See note (4): there are 433 MARRIED respondents which comprise 40.5% of the total respondents. See note (5): the total number of respondents is 1,068.
COLUMN TOTAL	The number of respondents in each column is listed across the bottom of the table.	See note (6). There are 453 respondents (42.4% of the total) who do have residences in their homeport.
COUNT	The number of people.	See notes (7). There are 327 MARRIED respondents with homeport residences.
ROW PCT	The COUNT expressed as a percentage of row total.	See notes (8). 327 people comprise 75.5% of the 433 MARRIED people.
COL PCT	The COUNT expressed as a percentage of column total.	See notes (9). 327 people comprise 72.2% of the 453 people with homeport residences.

LABEL	DESCRIPTION	EXAMPLE
TOT PCT	The COUNT expressed as a percentage of the total.	See notes (10). 327 people comprise 30.6% of the 1,068 total people.

Pages 8 and 9 of the printout are read the same as page 7, only these pages involve three units of information rather than just one. When comparing three units of information, a separate two-way table is generated for each value of the third unit of information. In the example given on pages 8 and 9, MARITAL STATUS is the third unit of information (see note (1) on page 8). Since MARITAL STATUS has 2 values (the value 1 for married and the value 2 for single) two tables have been generated.

On page 10, there is an illustration of the flexibility of CROSSTABS. Notice that the COUNT is the only statistic that is printed. Similarly, tables can be generated which contain the COUNT and any combination of the statistics ROW PCT, COL PCT, and TOT PCT.

In addition to the statistics appearing in each cell of the CROSSTABS tables, statistics may be requested which will be printed beneath the tables. These include Chi square, Phi or Cramer's V, contingency coefficient, Lambda, uncertainty coefficient, Kendall's tau b, Kendall's tau c, Gamma, Somer's D and Eta.

BREAKDOWN
(includes the CROSSBREAK facility)

This SPSS program generates statistics such as sums of answers and average answers for a unit of information which is separated into subgroups. The BREAKDOWN program should be used when a sum or an average figure is desired for an entire group of answers and also for various subsets of that group, such as the total number, or sum, of people in GAO and also sums for number of people in each division in GAO.

Examples of BREAKDOWN are given on pages 11 through 15 of the appendix. The information used to produce this output came from several sources. The first page of output was gathered by distributing a questionnaire to 159 Branch or Division Heads in Navy Headquarters organizations in the Pacific Fleet. The responses to the questionnaire determined present staffing of and duties performed by the organizations. Pages 12 through 15 were derived from the questionnaire described in the CONDESCRIPTIVE section.

There are three basic formats for the output. The first is a columnar table as shown on pages 11 and 12 of the appendix. The tops of these pages are read as follows (refer to page 12 of the appendix):

STATISTIC	DESCRIPTION	EXAMPLE
CRITERION VARIABLE	The coded name and description of the information from which the statistics are computed.	See note (1). The statistics refer to MCOST, COST OF PERSONNEL.
BROKEN DOWN BY	Introduces the coded name and description of the information whose values define the first set of subdivisions of the whole group.	See note (2). MCOST is divided into subdivisions by the values of Q2, MARITAL STATUS.
BY	Introduces the coded name and description of the information whose values form the second set of subdivisions of the whole group. (If there are more subdivisions the information involved would be listed following this section.)	See note (3). Each division by MARITAL STATUS is subdivided by the values of Q1, PAY GRADE.

The example on page 11 gives statistics for the whole group (see note (1) on page 11) followed by the statistics for 5 subdivisions of the group (see note (2) on page 11). The second example, on page 12, gives statistics for:

1. The whole group - see note (4) on page 12.

2. Three subdivisions of the group - see notes (5) on page 12.

3. Up to nine subdivisions of each of the three original subdivisions - see notes (6) on page 12.

The columns of the table are interpreted as follows. Refer to page 12 for examples.

STATISTIC	DESCRIPTION	EXAMPLE
VARIABLE	The coded names of the information forming the subdivisions.	See note (7). The subdivisions are formed by values of Q2 and then by values of Q1.
CODE	The value of the information.	See notes (8). Q2 has values 1, 2, and 3.
VALUE LABEL	The description of the values of the units of information.	See note (9). The value of 1 to Q2 corresponds to an answer of MARRIED to MARITAL STATUS.
SUM	The sum of all the answers given.	See note (10). The total cost of personnel for the whole group is $9,111,796.90 (the numbers are in thousands of dollars).
MEAN	The sum of all the answers divided by the number of answers; i.e., the average answer.	See note (11). The average cost of personnel for married respondents is $9,862. See note (12). The average cost of married personnel at grade level 1 is $5,782.
VARIANCE	A measure of how close the answers are to the mean.	See note (14). The variance (.002) is small relative to the mean (13.560) so there is almost no divergence in this subset of values.

STATISTIC	DESCRIPTION	EXAMPLE
STD DEV	The standard deviation is a measure of how close the answers are to the mean answer. This statistic is in the same units as the mean. STD DEV = $\sqrt{\text{VARIANCE}}$.	See note (13). The standard deviation (2.2054) is fairly small relative to the mean (8.5798) so there is not much dispersion of the values to MCOST.
N	The number of cases analyzed. This number does not include MISSING CASES (see interpretation below).	See note (15). There were 1,062 respondents in all. See note (16). There are 32 respondents who are grade level 2 and are MARRIED.
TOTAL CASES	The total number of responses.	See note (17) (bottom of page). There were 1,064 responses, but not all of these were analyzed.
MISSING CASES	The number of cases in which there were answers which were not analyzed. These are designated by the programmer.	See note (18). There were 2 respondents whose answers were deleted.

The second format of output is shown on page 13 of the appendix. The information in this format is interpreted the same as the first type of output, only it is arranged in a modified tree diagram rather than in a table.

The third format of output is requested by the CROSS-BREAK facility of BREAKDOWN. Pages 14 and 15 of the appendix illustrate this format. Notice that the statistics are printed in a table like the CROSSTABS format. Each cell of the table contains MEAN, COUNT, SUM, and STD DEV (see note (1) on page 14). The COUNT is the number of responses falling in the cell; e.g. (see note (2) on page 14), there are 142 respondents who are very satisfied with their job's use of their experience and are trained for their current job. The other statistics are interpreted the same as in the first format. Notice on page 15 that, as with CROSSTABS, any combination of the four statistics can be requested in the table. This example gives only the MEAN and COUNT. The CROSSBREAK facility can only be used with information whose values are whole numbers.

In addition to the statistics appearing in each cell of the CROSSBREAK tables, statistics may be requested which will be printed beneath the tables. These include Chi square, Phi or Cramer's V, contingency coefficient, Lambda, uncertainty coefficient, Kendall's tau b, Kendall's tau c, Gamma, Somer's D, and Eta.

T-TEST

This SPSS program generates statistics which compare either (1) two subdivisions of one unit of information or (2) paired answers for two units of information. This program is used to test whether or not the means of the two groups are the same. Explanations for the two types of output follow.

First, the example on page 16 of the appendix shows output for the comparison of two subdivisions of one unit of information. These two subdivisions are defined and the difference between their means is computed and evaluated. The program first assumes that the dispersion of the values for each subdivision is small, or that the values are close to the mean value. An F statistic, or test of significance, and its probability are computed. If the probability for F is low (i.e., on the order of 0.050 or lower), the assumption of equality of variances is rejected, and the reader should use the statistics under the heading SEPARATE VARIANCE ESTIMATE. If the probability of F is high use the section entitled POOLED VARIANCE ESTIMATE. Next, the assumption is made that the means or average values of the unit of information tested are equal for the two subdivisions. Then a T statistic, or test of significance of the assumption, is computed along with its probability. If the probability of the T value is low (again on the order of 0.050 or lower) then the assumption of equal means is rejected; i.e., the subdivisions have significantly different means, so they are dissimilar. However, a high probability of the T value indicates that the subdivisions are very similar.

For an illustration, refer to page 16 of the appendix. For this output, questionnaire responses from crew members on ships undergoing overhaul were used. (See CONDESCRIPTIVE for more details.) A subset of 1,062 responses was selected for analysis. The subdivisions of the data are defined (see note (1)). Within the dashed lines, the test information is shown by its coded name and description (see note (2)). In this case MCOST, COST OF PERSONNEL is tested. The following statistics are given for each subdivision.

STATISTIC	DESCRIPTION	EXAMPLE
NUMBER OF CASES	The number of responses in each group or subdivision.	See note (3). There are 631 answers in Group 1 and 431 answers in Group 2.
MEAN	The sum of the values divided by the number of values; i.e., the average value.	See note (4). Average COST OF PERSONNEL for Group 1 is $7,703 (the numbers are in thousands of dollars).
STANDARD DEVIATION	A measure of how close the answers are to the mean.	See note (5). The standard deviation for Group 2 is small relative to the mean so there is little dispersion of the values.
STANDARD ERROR	A measure of the accuracy of the value of the mean.	See note (6). The mean for Group 1 is fairly accurate since the standard error (.058) is small.

In the section following the general statistics, the F value is computed (see note (7)). It has a low probability as shown (0.000); therefore, the variations of values from the mean are significant, so the section entitled SEPARATE VARIANCE ESTIMATE is used. The T value also has a low probability (see note (8)) which indicates that the subdivisions are significantly different from one another.

The example on page 17 of the appendix shows the second type of output, comparing values of two different units of information. This data came from the same 1,062 cases previously mentioned. These two units of information are defined and statistics on each group are computed. Then the output gives statistics on the difference between the means of the two units of information. The program assumes that the difference between means is zero. Finally, a T statistic, or test of significance, and its probability are computed. If the probability of T is low, then the difference between means is greater than zero and the two groups are dissimilar. On the other hand, a high probability of T indicates that the difference between means is virtually zero and the two groups are very similar.

For an illustration, refer to page 17 of the appendix. The two units of information are defined (see note (1)) by their coded names and descriptions. Once again, NUMBER OF CASES, MEAN, STANDARD DEVIATION, and STANDARD ERROR are given (see previous explanation of these statistics). The next five columns are interpreted as follows. Refer to page 17 of the appendix for examples.

STATISTIC	DESCRIPTION	EXAMPLE
(DIFFERENCE) MEAN	The value of the difference between the means of the two units of information.	See note (2). MEAN for Q2 minus MEAN for MCOST = 1.6610 - 8.5791 = -6.9181.
STANDARD DEVIATION	This measures how closely the differences between means fall around the average difference between means.	See note (3). Standard deviation is 2.514 and is small relative to the average (-6.9181) so there is little dispersion of values.
STANDARD ERROR	This indicates how accurate the average difference is.	See note (4). The small standard error (.077) implies the mean is accurate.
CORR.	This is an indication of the relationship between the two units of information; it is called a correlation coefficient. For a more detailed explanation of correlation, see the chapter on PEARSON CORR.	See note (5). The correlation is -.414 which indicates that the two units of information are significantly related as shown by the two-tailed probability.
2-TAIL PROB.	This is a measure of the likelihood of getting the above correlation coefficient from a random sample.	See note (6). The low probability indicates that this correlation is significant.

In this example the T value has a very low probability (see note (7)); therefore, the difference between means does not equal zero and the two units of information form significantly different groups.

PEARSON CORR

This SPSS program generates statistics which describe how well one unit of information can be used to estimate another unit of information. These statistics define a relationship known as the correlation between two units of information. The PEARSON CORR program measures fluctuations in the two units of information to see if there is any consistency in the changes, for example, whether pay grade increases as salary increases; whether age increases as salary increases, decreases, or remains unchanged; and whether turnover rates decrease as unemployment increases. It is important to remember that correlations indicate relationships but do not indicate causality. For example, age and occurrence of heart attacks may be correlated; however, age itself does not cause the heart attack.

An example of the output from PEARSON CORR is on page 18 of the appendix. Once again, for this page of output cases were selected from the questionnaire distributed to crews on ships undergoing overhaul. The coded names and descriptions of the units of information being compared are listed at the top and to the left of the statistics. The three numbers on page 18 of the appendix are the correlation coefficient, the number of answers analyzed, and the significance of the correlation coefficient. These are explained in the following chart (see page 18 of the appendix for references). Note that the bottom line on the printout (see note (6) on page 18) is a reminder of what the numbers represent.

STATISTICS	DESCRIPTION	EXAMPLE
Q1	The coded name of one of the units of information being compared.	See note (1). Q1 represents pay grade.
MCOST	The coded name of the other unit of information being compared.	See note (2). MCOST represents salary.
Correlation Coefficient	This number is a measure of the strength of association between two units of information and is usually represented as r. Numbers close to ± 1.0 indicate strong relationships. Numbers close to 0.0 indicate very weak or no relationships. Positive numbers indicate that as one unit of information increases or decreases, the other unit of information changes in the same direction. Negative numbers indicate that as one unit of information increases or decreases, the other unit of information changes in the opposite direction.	See note (3). r = .9709 indicates that as pay grade increases, salary increases and the two are very good predictors of one another.
Number of Cases	This is the number of cases used in the analysis, usually denoted as N. This number excludes cases which contained information that was declared "MISSING" or invalid by the programmer.	See note (4). N = 1,063.
Significance of r	This indicates the statistical significance of the correlation coefficient, usually denoted as S. Low numbers indicate high significance. Acceptable significance varies depending on the number of cases being analyzed.	See note (5). S = .001 indicates that the correlation coefficient is highly significant.