CDC 2R051

Maintenance Management Analysis Journeyman

Volume 3. Statistical Methods

Air Force Career Development Academy

Air University

Air Education and Training Command

2R051 03 1705, Edit Code 07

AFSC 2R051

Author: MSgt Clifton Yelverton
363rd Training Squadron
82nd Training Group (AETC)
363 TRS/TRR
520 Missile Road
Sheppard Air Force Base, Texas 76311–2261
DSN: 736–2146
E-mail address: clifton.yelverton@us.af.mil or
363TRSCDCWriters@us.af.mil

Instructional Systems Specialist: Evangeline K. Walmsley

Editor: Catherine Parker

Air Force Career Development Academy (AFCDA)
Air University (AETC)
Maxwell-Gunter Air Force Base, Alabama 36118–5643

Material in this volume is reviewed annually for technical accuracy, adequacy, and currency. For SKT purposes the examinee should check the *Enlisted Promotions References and Requirements Catalog* to determine the correct references to study.

THIS CAREER DEVELOPMENT course, *2R051, Maintenance Management Analysis Journeyman,* volume 3 teaches the statistics principles and applications that every maintenance management analysis journeyman needs to know and be able to apply to become effective in your job. It is presented in four units.

Unit 1 begins with a review of several mathematics fundamentals and introduces you to statistical methods and operations. This unit will prepare you for units 2, 3, and 4.

Unit 2 covers descriptive statistics that include the measurement scales, data distribution and its graphical representation, the measures of central tendency and a description of the different kinds of data distribution curves. It also explains sampling theory, and how to measure the variability of data from the mean. You will learn important concepts such as frequency distributions, standard deviations, and standard errors. We will cover the concept of hypothesis testing along with developing a hypothesis and the procedures for testing a hypothesis.

Unit 3 deals with statistical process control. The process of selecting, constructing, and interpreting control charts is presented.

Unit 4 covers predictive analysis, a major work of the analyst. We cover correlation, trend analysis, extrapolation, regression, and probability. Here, you learn to measure the relationship between two sets of data.

After completing this volume, you should have a sound basis for performing elementary statistical analysis. The knowledge you gain from this volume will give you the foundation for a sound analytical approach to problem solving in the maintenance process. As an analyst, you should increase your skill and knowledge of statistics by studying additional statistics sources (e.g., journals, Internet websites, etc.). Do not base your statistical work solely on the knowledge obtained from this volume.

A glossary is included for your use.

Code numbers on figures are for preparing agency identification only.

The use of a name of any specific manufacturer, commercial product, commodity, or service in this publication does not imply endorsement by the Air Force.

To get a response to your questions concerning subject matter in this course, or to point out technical errors in the text, unit review exercises, or course examination, call or write the author using the contact information on the inside front cover of this volume.

NOTE: Do not use Air Force Instruction (AFI) 38−402, *Airmen Powered by Innovation*, to submit corrections for printing or typographical errors. For Air National Guard (ANG) members, do not use Air National Guard Instruction (ANGI) 38−401, *Suggestion Program.*

If you have questions that your supervisor, training manager, or education/training office cannot answer regarding course enrollment, course material, or administrative issues, please contact Air University Educational Support Services at http://www.aueducationsupport.com. Be sure your request includes your name, the last four digits of your social security number, address, and course/volume number.

This volume is valued at 21 hours and 7 points.

NOTE:

In this volume, the subject matter is divided into self-contained units. A unit menu begins each unit, identifying the lesson headings and numbers. After reading the unit menu page and unit introduction, study the section, answer the self-test questions, and compare your answers with those given at the end of the unit. Then complete the unit review exercises.

Page

Unit 1. Fundamentals of Mathematics

MUCH OF YOUR WORK as an analyst requires you to have a sound knowledge of mathematics and statistics. With this knowledge, you can state facts, identify potential problems, draw preliminary conclusions, or make predictions about the data involved. Statistics is not something new to us; it is a part of our everyday lives and thinking processes. When we say something is typical, normal, average, or abnormal, we are thinking in statistical terms. Therefore, statistics is not foreign; it is a scientific form to our everyday thinking. As you may already realize, you will probably use the information in this unit daily; therefore, make every effort to retain as much of it as possible.

This unit discusses certain mathematical concepts as they apply to statistics. You will also learn two major, statistical methods and operations used in statistics.

1−1. Review of Mathematical Concepts

Many of the errors that occur in statistical computations are not caused by insufficient knowledge of statistical procedures but are caused by mistakes in simple mathematics. To understand how to perform statistical computations effectively, you need to understand some basic mathematical concepts. In this section, we discuss natural and signed numbers, and rounding numbers. Let’s start this section with a review of natural numbers.

401. Review of arithmetic

Statistics is about numbers. It is important to know the basic properties and characteristics of numbers. It is equally important to know how to use arithmetic operations to get the right results. This lesson briefly reviews elementary and high school math. These concepts and principles will serve as the basic math structure for most of your statistics equations.

Natural numbers

A *natural number* is simply a positive number (+) that has any actual value from zero (0) to infinity (∞). The actual value of a number is written with its sign included, for example, +6. However, normally a positive value, such as +6, is written with the plus sign understood (simply 6). With this understanding of what constitutes the value of a number, consider the four basic mathematical operations—addition, subtraction, multiplication, and division.

Each mathematical operation has its own unique properties. You must understand each operation separately in order to use them collectively when solving complex formulas. Note the terminology associated with each operation.

1. When numbers are added, the answer is called a *sum*.

   Example:

   4 + 2 = 6 (sum)

2. When numbers are subtracted, the answer is called a *difference*.

   Example:

   5 – 2 = 3 (difference)

3. When numbers are divided, the answer is called a *quotient*.

   Example:

   10 ÷ 2 = 5 (quotient)

4. When numbers are multiplied, the answer is called a *product*.

   Example:

   3 × 3 = 9 (product)

Order of operation

When using two or more of these operations together, you must know their order of operation. Your answer will be *incorrect* if the operations are *not* performed in the *proper order*. The example below shows the correct way to solve a problem having multiple operations and it also shows how an incorrect answer is attained if mathematical functions are improperly grouped.

Grouping

To avoid problems caused by incorrect order of operation, mathematicians developed a shorthand notation method. Symbols are used to indicate how numbers should be combined and used with the various operations when solving mathematical problems. Symbols such as parentheses (), brackets [], and braces {}, are used as symbols of grouping when more than one operation is used in a mathematical problem. You must solve the equation enclosed in parentheses, brackets, and braces before solving other portions of the problem. In situations where symbols are used, you should work from the inside out and in the order of parenthesis, brackets, and braces. The symbols of grouping are used to indicate the order in which the operations of multiplication, division, addition, and subtraction must be performed when solving problems. The acronym, PEMBAS, is often used to help remember the order of the operation: parenthesis, exponents, multiplication and division, and finally addition and subtraction or PEMDAS.

For example, instead of writing (9 – 6 ÷ 3), write [9 – (6 ÷ 3)]. The latter equals [9 – 2] or 7. These symbols of grouping indicate you must divide before subtracting. Listed next are four basic rules that you must follow when solving mathematical problems.

Rule 1

A general rule to follow concerning the *order* in which operations are performed (provided no symbols of grouping are involved) is to *multiply and/or divide from left to right, before adding and/or subtracting from left to right*.

Examples:

8 – 6 + 3 × 6 + 8 ÷ 2 = (compute 3 × 6 and 8 ÷ 2 first)

8 – 6 + 18 + 4 = (compute 8 – 6 next)

2 + 18 + 4 =

20 + 4 = 24

4 × 3 ÷ 2 – 5 + 2 = (compute 4 × 3 first)

12 ÷ 2 – 5 + 2 = (compute 12 ÷ 2 next)

6 – 5 + 2 = (compute 6 – 5 next)

1 + 2 = 3

Rule 2

Consider numbers above or below a fraction line as single numbers, and *combine* them before dividing.

Example:

$$\frac{5+3}{8-4}=\frac{8}{4}=2$$

Rule 3

A number either placed directly in front of or directly behind parentheses, brackets, or braces, with no sign of operation between, indicates that the quantity within the parentheses, brackets, or braces must be multiplied by that number.

Example:

$9 - 3 + 2(6+2) =$

$9 - 3 + 2(8) =$

$9 - 3 + 16 =$

$6 + 16 = 22$

Rule 4

If a symbol of grouping encloses one or more additional symbols of grouping, remove them by removing the *innermost symbol first*, and so forth, until the last one is removed.

Example:

$3[4 + (3 \times 2\ \{6 \div 2\})] =$

$3[4 + (3 \times 2\ \{3\})] =$

$3[4 + (3 \times 6)] =$

$3[4 + 18] =$

$3[22] = 66$

Signed numbers

The concept of signed numbers involves values that are positive (+) and those values that are negative (–). As stated earlier, positive numbers may be written without the (+) sign. However, you must *always* write the negative or minus sign (–) with negative numbers (–6 for example) because the negative sign is *never* assumed. In addition to the rules associated with using more than one mathematical operation, as stated earlier, negative numbers also have specific rules for operations.

Addition

To add a series of numbers, each of which is negative, add the absolute values in the usual fashion and place a minus sign in front of the answer.

Example:

$(-7) + (-3) + (-2) = -12$

To add a series of numbers with mixed signs, add all the positive numbers and then add all the negative numbers. Next, find the difference between the two totals, and give the answer the sign of the larger value.

Example:

(+4) + (–7) + (+12) + (–4) = (+16) + (–11) = +5

If there are only two numbers, one negative and the other positive, subtract the smaller from the larger and give the answer the sign of the larger value.

Example:

(–14) + (+8) = –6

Subtraction

To subtract a negative number, change its sign and proceed as in addition.

Examples:

(+6) – (–3) = (+6) + (+3) = +9

(–5) – (–8) = (–5) + (+8) = +3

Notice in the example, the operational sign is changed to positive at the same time the sign of the subtrahend (a quantity or number to be subtracted from another) is changed. This is because you are now adding.

Multiplication

When two negative numbers are multiplied, the product of the two numbers is positive. But, when a negative number and a positive number are multiplied, the product of the two numbers is negative.

Examples:

(–5) × (–3) = +15

(–2) × (+3) = –6

Division

When a negative number is divided by another negative number, the answer is positive. When the dividend and the divisor have unlike signs, the answer is negative.

Example:

$$\frac{-12}{-6} = 2 \quad \frac{-12}{6} = -2 \quad \frac{12}{-6} = -2$$

Remember, in division, *dividing by zero is a prohibited operation*. The result of such attempt would be undetermined.

Rounding numbers

Until now, all mathematical problems were carefully selected to avoid decimals. In actuality, however, most of your calculations will involve decimals. For that reason, it is important that you round your answers to the correct number of digits. Also, as you will learn when performing advanced statistical formulas, decimals will determine if potential problem areas require investigation.

Accuracy of digits

In a course involving statistical computations, the question of how many decimal places to carry is almost invariably asked. There is no simple answer to this question. The accuracy of results obtained by statistical computations depends upon both the accuracy of the original data and the computations to which the data are subjected. One answer rounded to the nearest whole number may still contain

some inaccurate digits, while another answer rounded to four decimal places may contain no inaccurate digits. In the absence of good rules, you could drop too many decimals and lose some of the accuracy that you really have.

On the other hand, you could save a string of decimals beyond the limits of accuracy, giving the appearance of great exactness that really does not exist. Common usage in some cases establishes the number of decimals required, but in many instances the decision rests with the person making the computations. The following discussion of the general rules for rounding numbers should make you more aware of the problems of accuracy in computations.

Rules for rounding

A good rule to begin with is to round *only* the final answer; do *not* round intermediate totals leading to the final answer.

To round numbers to the nearest whole number or to the nearest decimal place, proceed as follows.

1. To the nearest *whole number*:

   6.2 becomes 6

   7.81 becomes 8

2. To the nearest *tenth*:

   5.173 becomes 5.2

   5.11 becomes 5.1

3. To the nearest *hundredth*:

   6.177 becomes 6.18

   0.574 becomes 0.57

   2.0982 becomes 2.10

When the number to be rounded falls exactly halfway between two values, round up.

Examples:

8.5 rounded to the nearest whole becomes 9

5.15 rounded to the nearest tenth becomes 5.2

3.165 rounded to the nearest hundredth becomes 3.17

When approximate numbers are added or subtracted, the answer is accurate only as far as the number having the smallest number of decimal places (for example, 5.27 + 6.3 + 18.963 + 7.41 = 37.9, which was rounded from 37.943). This is illustrated below:

$$\begin{array}{r} 5.27 \\ 6.3 \\ 18.963 \\ +\underline{7.41} \\ 37.943 \cong 37.9 \end{array}$$

When two approximate numbers are multiplied or divided, the answer generally has no more accurate digits than the figure with the smaller number of accurate digits (for example, $23.57 \times 1.2 = 28$ rounded from 28.284). In other words, the apparent answer of 28.284 is probably accurate to only two figures. As another example, $6.5583 \div 2.1 = 3.1$ rounded from 3.123. Again, the apparent answer of

3.123 probably has only two accurate figures. The above rule also applies to the squares and square roots of approximate numbers, because squaring is a special case of multiplication, and finding the square root is a special case of division.

The above rules are not required when working with exact numbers. The product of two exact numbers is accurate to all obtained figures. Also, the quotient of two exact numbers may be carried to as many decimal places as desired.

The above general rules are just guidelines. Although the rules are valid most of the time, do not become a slave to them. In computations involving several steps, it is better to carry one more decimal place than would be required for strict accuracy under the rules. At the end of the solution, you can decide upon the extent of the accuracy in the answer by applying the rules to the various steps. Apply the general rules and use your experience to guide you in the rounding off process.

In this text, carry your decimals to the number of digits shown in the examples. Some of your answers may still differ slightly. Consider your answer correct if it is within ± .1 decimal.

Root numbers

The root of a number is one of two or more equal numbers which, when multiplied together, will produce the original number. Such a number is called an *equal factor*. Two equal factors that will produce 9 is 3. The two equal factors are called *square roots*. Therefore, the square root of 9 is 3. This may be written in mathematical expression as $\sqrt{9} = 3$. The symbol $\sqrt{\ }$ is the square root symbol. It is also called a *radical* sign. The following rule may be helpful when solving formulas using root numbers: To obtain the square root of a fraction, find the square root of the numerator and denominator then divide:

$$\sqrt{\frac{100}{25}} = \frac{\sqrt{100}}{\sqrt{25}} = \frac{10}{5} = \mathbf{2}$$

402. Computing percentages

As a maintenance analyst, you will frequently solve percentage-type problems. Figures expressed as percentages are used for comparison because they are easily understood. Often, the effectiveness or efficiency of equipment and personnel is found by using the percentage formula. They are extremely useful when reporting equipment status, computing man-hour utilization, and projecting capabilities.

Percentage

The term percent is symbolized by the percent sign (%). When you express figures in percent, you are expressing them in terms of “how many per hundred.”

To change a decimal to a percent, move the decimal point two places to the right and add the percent sign. Inversely, to change a percent to a decimal, move the decimal point two places to the left and drop the percent sign.

Example:

.635 = 63.5%

.0035 = .35%

.8 = 80% or 80.0%

Each percentage problem consists of the following components:

Rate (R): The number with either % or the word “percent” following it. The rate is the number that denotes how many hundredths are taken.

Portion (P): The part of the whole in the problem.

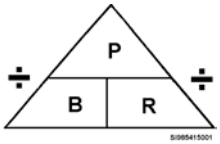
Base (B): The quantity of which the percent is taken. It is the whole upon which the problem is based.

Calculating unknown values

The formula used for computing percentages is as follows:

$$\frac{P}{BR} = 1$$

Figure 1–1 shows the relationship of the three components by means of a pyramid equation. By substituting any two of the known values into the formula, you can solve for the unknown. The problems that follow demonstrate the types of percentage problems you will solve. Solutions to the problems are included.



$P = (B)\ (R)$

$B = \frac{P}{R}$

$R = \frac{P}{B}$

Figure 1–1. Proportion, rate, and base pyramid.

Finding the rate

To find the rate, simply divide the portion by the base to get a decimal. Then change this decimal into the percent form.

Example:

23 is what percent of 41?

$B = 41\ R = \frac{P}{B}$

$P = 23\ R = \frac{23}{41}$

R = ? R = .561 or 56%

Finding the portion or part

To find P, follow the formula and multiply the base (B) times the decimal equivalent of the rate (R).

Example:

What is 15 percent of 95?

B = 95 P = (B)(R)

P = ? P = (95)(.15)

R = 15% P = 14.25

Finding the base

To find the base (B), follow the formula, change the rate to a decimal, and divide the portion by this decimal.

Example:

150 is 70 percent of what number?

$B = ?\ B = \frac{P}{R}$

$P = 150\ B = \frac{150}{.70}$

$R = 70\%\ B = 214.29$

The pyramid equation is a quick reference formula to use when solving for the rate, portion, or base. If any of the two variables are known, you can calculate the other by using the formula.

Self-Test Questions

After you complete these questions, you may check your answers at the end of the unit.

401. Review of arithmetic

1. What are natural numbers?

2. What is the general rule of operations when working with a grouping of numbers with no symbols of grouping involved?

3. What are signed numbers?

4. How do you add two numbers with different signs?

5. What is the result when one negative number is divided by another negative number?

6. What is a good rule to follow when rounding?

7. Round the following numbers to the nearest whole number:

   a. 7.43.

   b. 8.64.

   c. 17.77.

8. Round the following numbers to the nearest tenth:

   a. 3.173.

   b. 7.428.

   c. 2.65.

9. Round the following numbers to the nearest hundredth:

   a. 8.121.

   b. 6.7379.

   c. 3.415.

10. Write this phrase as a mathematical expression: “The square root of 16 is 4.”

402. Computing percentages

1. How is the rate determined?

2. How is the portion determined?

3. How is the base determined?

1−2. Statistical Methods and Operation

This section discusses the two major divisions of statistical methods—descriptive and inferential statistics. It also includes the basics of statistical operation to enable you to understand and perform your computations.

403. Understanding statistical methods

Statistics deals with scientific methods for collecting, organizing, analyzing, summarizing, and presenting data, as well as drawing valid conclusions and making reasonable decisions based on the analysis. It is applied to many different problems so it is important to understand it from the viewpoint of application and interpretation.

The performance of one group of data can be compared with another, and the significance of difference can be tested. Suppose, for example, you are comparing the performance of two aircraft to determine which one will be used in an upcoming exercise. After collecting sufficient data, you may recommend an aircraft to perform this mission.

The process of identifying, organizing and presenting data represents descriptive statistics; testing, analysis, and predicting are forms of inferential statistics. Descriptive statistics is an objective approach, whereas inferential statistics is a subjective approach.

Descriptive statistics

Descriptive statistics describe the actual data being observed. You collect and summarize data without drawing any conclusions or inferences. The monthly maintenance summary is an example of descriptive statistics. It involves a study of the characteristics of the items (data) being observed. Descriptive statistics is used to *summarize* large amounts of data. As you will see, descriptive statistics provide a number of useful ways of depicting a mass of numbers so they become more meaningful and more readily conveyed to others. You describe and measure a given amount of data without drawing any conclusion or inferences about the data.

Descriptive statistics includes the identification and classification of data. You will learn the different types of data and the measurement scales used to describe them. Organizing data involves the use of frequency distribution to group data. Using graphs and diagrams to represent data gives a better picture of the data distribution when presenting the collected data. When you measure data, you determine the measure of central tendency. Variability describes the differences between data in a normal distribution. Correlation and regression can also be used to describe data.

Inferential statistics

Inferential statistics involves sampling parts of the data and then *drawing conclusions* or *generalizations* about the data based on those samples. Inferential statistics are statistical techniques or methods that allow you to make inferences. That is, they let you generalize from the results of small samples of subjects to those of the larger groups they represent. Although the inferential method uses *only a sample of data*, the sample must be representative of the data being studied. For example, you cannot use only January's data as a sample when studying an entire year's data. It is also important that your representative sample be chosen at random. This means that each item of data must have an equal chance of being included in the sample. In other words, a 12-month study must have samples from each month rather than all samples from only one month.

The process of inferential statistics begins when you analyze data to draw some conclusions about the data. Statistical process control involves the use of standards or control to measure the quality of data. Hypothesis testing involves sampling, analyzing, and testing the validity of data to qualify or disqualify an assumption about the data. Predictive analysis uses various methods to make predictions. These methods are correlation and regression analysis, trend analysis, extrapolation, and probability.

404. Performing statistical operations

In your brief career as an analyst, you undoubtedly have seen statistical problems containing symbols. Symbols are actually a form of shorthand that allows you to say a lot with little effort.

To do an effective statistical study on maintenance data, you need to understand how statistical symbols are used. In this lesson, we discuss the various statistical operation and symbols you need to know when doing your analytical studies involving statistics.

Variables

A *variable* is a symbol that can assume any of a prescribed set of values. You usually use letters, such as A, a, K, X, N, or Greek symbols, such as σ, α, or β, to represent variables. The most common letters used in statistics to represent variables are X, Y, and N; each letter denoting a set of values. However, you can use any letter to represent a set as long as you define what set the letter represents. For example, you can say that X represents all aircrew personnel, Y represents maintenance personnel, and N represents the combination of all aircrew and maintenance personnel. You can use H to represent the height of all personnel. If the variable can assume only one value, then it is called a constant. The letter K, or k, is used most of the time to symbolize a constant.

Equations

Equations are statements in the form A=B ("*Variable* A is equal to *variable* B" or "A equals B"). The *variable* A is on the left-hand side of the equation while the *variable* B is on the right-hand side. The value of each side of the equation must equal each other; the values of the variables, however, are not necessarily the same.

For example, using the equation A=B, when A=1, B=1; A=2, B=2; A=5, B=5.

If the equation is A+1=B, then if A=1, B=2; A=2, B=3, A=5, B=6.

Functions

When two or more sets of values are related to each other, then a *function* between the values exists. Let's say the variable X represents a value in a set of values called X-values, and the variable Y represents the corresponding values in a set called Y-values. When the values of Y depend on the values of X, then you say that Y is a *function* of X. This is written in the form Y = F(X) (read "Y equals F of X"). The variable X is the independent variable and the variable Y is the dependent variable.

For example, if the speed *s* of the aircraft depends upon the time *t*, then you write *s* = F(*t*)n, which means that the speed of the aircraft is a function of time.

Throughout the study of statistics you will encounter functions and equations and sets of values. We use a *table* to depict the sets of values used, whether assumed or computed, to follow the order of values in an equation.

Below is the result of equation A+1=B, where B is a function of A or B=F(A).

When A is:	1	2	3	4	5
B is equal to:	2	3	4	5	6

Subscripts

Small numbers or letters written slightly below and to the right of a symbol are called *subscripts*. Subscripts are used to denote the different individual values in a set of data, to indicate which set of data a certain measure represents, to denote the different individual points in a series, and so forth. For example, the different individual values in a series of X values may be denoted by numeric subscripts in the following manner: X_1, X_2, X_3, X_4, ..., X_n. They are read as X sub-one, X sub-two, X sub-three, and so forth. As another example, S is used as a symbol for the sample standard

deviation of a set of data. As long as you are working with only one set of data, a subscript is not needed. But what if you are computing the standard deviation for two separate sets of data—a set of X values and a set of Y values? You can add a subscript to the S and use S_x (S sub-x) to represent the standard deviation of the X values and Sy (S sub-y) to represent the standard deviation of the Y values.

Summations

One of the most widely used statistical symbols is the summation sign (Σ). Σ is the capital Greek letter sigma and is read as "the sum or summation of." The summation sign directs you to sum (add) all the individual values in a series.

Example:

$\Sigma X = X_1 + X_2 + X_3 ... X_n$

There are occasions when a more precise way to indicate summations is needed to direct summation of only part of a series. However, in this course the summation sign means to sum all the values in the series. You will also be using Arabic letters as symbols for statistics. In the remaining discussion, Greek letters are used as symbols for parameters. For example, the mean of a population is represented by the Greek letter mu (μ), whereas the mean of a sample taken from that population is represented by $\overline{X}$. The standard deviation of the population is represented by the lower case Greek letter sigma (σ); however, the standard deviation of the sample is represented by s.

Exponents

When one number, such as the base, is used as a factor two or more times, the result is a power of the base, which is called an *exponent*. A positive integer or exponent written as a small number just to the right and slightly above the base number indicates the number of times the base number is used as a factor. In the value 4^2, for example, 4 is the base and 2 is the exponent. The value 4^2 may be read as follows: 4 squared, 4 to the second power, or 4 superscript 2. Thus, the equation 4^2 means 4×4, which is 16. As another example, you solve 6^3 as follows: $6 \times 6 \times 6 = 216$. The following rules will help you solve formulas using exponents:

1. When multiplying two powers of the same base, add the exponents:

   $a^3 \times a^2 = (a \times a \times a)(a \times a) = a^{3+2} = a^5$

2. When obtaining a power of a power, multiply the exponents:

   $(a^2)^3 = (a^2)(a^2)(a^2) = (a \times a)(a \times a)(a \times a) = a^6$

3. When dividing one power of a specified base by another power of the same base, subtract the exponent:

   $$\frac{a^4}{a^2} = a^{4-2} = a^2$$

4. To obtain a power of a product, raise each factor of the product to the specified power and multiply:

   $(abc)^2 = a^2 \times b^2 \times c^2$

5. To obtain a power of a fraction, raise the numerator and the denominator to the specified power and divide:

$$\left(\frac{a}{a}\right)^4 = \frac{a^4}{a^4}$$

6. Every number is equal to its own first power. Numbers that are divided by their own value are equal to 1. Therefore, any number raised to the power of zero is equal to 1.

$$\frac{a^4}{a^4} = a^{4-4} = a^0 = 1$$

7. When solving equations with negative exponents, remember the laws of exponents for multiplication:

$a^3 \times a^{-3} = a^{3-3} = a^0 = 1$

Then, if a^3 times a^{-3} is equal to 1, you can rearrange the formula and say that $1 \div a^3$ must be equal to a^{-3}. Hence $a^{-3} = 1 \div a^3$, or for any positive value, $a^{-n} = 1 \div a^n$.

When solving formulas that use exponents, remember the bases of the exponents must be identical in order to combine them.

Thus, $a^5 \times b^5$ cannot be combined because the bases (a and b) are not the same. Additionally, no number may be divided by zero; this is a prohibited calculation. Some examples of simple problems involving exponents are shown below:

$$(4^2)(2^3) = (16)(8) = 128$$

$$(5^0)(3^2) = (1)(9) = 9$$

$$4^3 \div 2^2 = 64 \div 4 = 16$$

Inequality

Five inequality symbols are used occasionally in the study of statistics. The symbol > means "is greater than," and the symbol < means "is less than." For example "N>30" is read "N is greater than 30," and "N<30" is read "N is less than 30." If you write "N≥30" it means "N is equal to or greater than 30," or if you write "N≤30," it means "N is equal to or less than 30." The symbol ≠ means "does not equal." It is the simple equals sign with a slash mark through it. It is used to indicate that two things are not equal. For example, "N≠30" is read "N does not equal 30."

The inequality symbols are as follows:

>	Greater than.
≥	Equal to or greater than.
<	Less than.
≤	Equal to or less than.
≠	Does not equal.

You may use these symbols to indicate a range of values.

For example, 2<N<10 is read as "N is greater than 2 and less than 10." This means that the variable N can be any number above 2 but below 10, which are the numbers 3, 4, 5, 6, 7, 8 and 9.

If you write it this way: $10 \geq N \geq 2$ ("N is equal to or less than 10 and equal or greater than 2"), this means that the values of N can be any number from 2 to 10.

NOTE: Additional symbols will be defined when they are first used in this text.

Self-Test Questions

After you complete these questions, you may check your answers at the end of the unit.

403. Statistical methods

1. What statistical method is considered an objective approach? Subjective approach?

2. Match each statistical method in column B with the statement in column A that best describes it. Each item in column B may be used once or more than once.

Column A	*Column B*
____ (1) Summarize large amounts of data.	a. Descriptive.
____ (2) Draws conclusions based on samples.	b. Inferential.
____ (3) Does not draw conclusions about the data.	
____ (4) Used only to describe data.	
____ (5) Involves testing the validity of data.	

404. Statistical operation

1. Which of the following are variables: 3, A, 4.5, +, Y?

2. In the function M=F(N), which one is the independent variable? The dependent variable?

3. Why are subscripts used?

4. What is the meaning of the symbol Σ?

5. What is the value of 2^5 equal to? Which one is the base? Which one is the exponent?

6. What is the range of the values of N in this expression "$5 \leq N < 15$"?

Answers to Self-Test Questions

401

1. Positive numbers (+) that have any actual value from zero (0) to infinity (∞).
2. Multiply and/or divide from left to right, before adding and/or subtracting from left to right.
3. Those values that are positive (+) and those values that are negative (–).
4. Subtract the smaller from the larger and give the answer the sign of the larger value.
5. Positive.
6. Round *only* the final answer; do not round intermediate totals leading to the final answer.
7. a. 7.
   b. 9.
   c. 18.
8. a. 3.2.
   b. 7.4.
   c. 2.7.
9. a. 8.12.
   b. 6.74.
   c. 3.42.
10. $\sqrt{16} = 4$.

402

1. Divide the portion by the base to get a decimal. Then change this decimal into the percent form.
2. Multiply the base (B) times the decimal equivalent of the rate (R).
3. Change the rate to a decimal and divide the portion by this decimal.

403

1. Descriptive. Inferential.
2. (1) a.
   (2) b.
   (3) a.
   (4) a.
   (5) b.

404

1. A, Y.
2. N, M.
3. To denote the different individual values in a set of data, to indicate which set of data a certain measure represents, and to denote the different individual points in a series, etc.
4. The summation sign, Σ, directs you to add all the values in a series.
5. 32; 2; 5.
6. 5 through 14.

Complete the unit review exercises before going to the next unit.

Unit Review Exercises

Note to Student: Consider all choices carefully, select the *best* answer to each question, and *circle* the corresponding letter. When you have completed all unit review exercises, transfer your answers to the Field-Scoring Answer Sheet.

Do not return your answer sheet to the Air Force Career Development Academy (AFCDA).

1. (401) A natural number is stated as a positive number that
   a. has any actual value from zero to infinity.
   b. has its own unique properties.
   c. can be added or subtracted.
   d. can be summed.

2. (401) Which phrase is true regarding signed numbers?
   a. Dividing by zero is permitted.
   b. Only whole numbers are used.
   c. Negative numbers are not used.
   d. Both positive and negative numbers have values.

3. (401) Compute the difference of the signed numbers –17 – (–10).
   a. +3.0.
   b. –3.0.
   c. –7.0.
   d. –27.0.

4. (401) When rounding numbers to the nearest tenth, what is the closest accurate sum of 6.5 + 7.31 + 8.433?
   a. 22.0.
   b. 22.2.
   c. 22.3.
   d. 22.24.

5. (401) How do you write: "The square root of 25 is 5." as a mathematical expression?
   a. $\sqrt{25 = 5}$.
   b. $25\sqrt{\ } = 5$.
   c. $\sqrt{\ 25 = 5}$.
   d. $\sqrt{25} = 5$.

6. (402) Eighty is 40 percent of what number?
   a. 200.
   b. 320.
   c. 400.
   d. 500.

7. (403) Which statistical technique is an example of descriptive statistics?
   a. Probability.
   b. Extrapolation.
   c. Trend analysis.
   d. Measurement scales.

8. (403) What statistical method uses *only a sample* of data?
   a. Inferential.
   b. Descriptive.
   c. Data survey.
   d. Randomization.

9. (404) In the equation Y = X + 2, what does the symbol "Y" represent?
   a. Exponent.
   b. Inequality.
   c. Subscript.
   d. Variable.

10. (404) What is the value of 4^3?
    a. 64.
    b. 24.
    c. 12.
    d. 7.

Student Notes

Unit 2. Descriptive Statistics

DESCRIPTIVE STATISTICS involves the use of numbers to summarize information already known about a given situation. In this unit, you will learn the various types of data collections known as data distributions. You will also learn about the means of summarizing these distributions by using measures of central tendency and variability.

2–1. Data Measurement

Since data is the foundation of any statistical study, it is most essential that the analyst recognize different types of data and handle them in the proper manner. This section discusses the identification and classification of data, how to sample data, and four different data measurement scales applicable to statistics. It is important to understand these concepts so you can perform accurate and reliable statistics for any data given.

405. Data and samples

Data is any thing or item known or assumed. It could be tangible objects (e.g., cars, people, food, etc.) or intangible information (e.g., numbers, intelligence quotient [IQ] etc.). In statistics, data must be measured or described, and organized. Data that can be described without any precise measuring unit are *nonquantitative data*, also known as *qualitative data.* Data that can be precisely counted or measured in terms of measuring units are called *quantitative data*. The color of one's eyes can be considered nonquantitative or qualitative data, so is gender (male or female). Height and length are examples of quantitative data because they are measured in inches, feet, or meters. In the world of maintenance data analysis, you will deal primarily with quantitative data, using facts and figures (numbers) to support a statistical study or presentation. So, whenever we mention the word "data" we refer to quantitative data, unless we say otherwise.

Types of data

In statistics, quantitative data can be classified as *continuous* or *discrete*.

Continuous data

Continuous data is data that can be measured with varying degrees of precision. For example, in measuring the length of an item in meters, you could break it down into centimeters and millimeters and, depending on the precision of your measuring device, into even smaller units. Such data can be considered points on a line. The data is considered continuous because it can continually be broken down into smaller and smaller units.

Discrete data

Discrete data is based on counting items that can be expressed only in whole number units. The number of technicians in an avionics shop can occur only in whole units. For example, if there are 7 technicians in the shop, that means 7, not 6.5 or 7.5 people. However, in statistics, most data tends to be treated as continuous, so it's not unusual to see statements such as, "On the average, it takes 3.4 technicians to replace an afterburner unit on a particular engine."

Sampling theory

Sampling theory is the study of the relationship between a sample and the larger group from which the sample was taken. Before doing a statistical study, you must collect sufficient data for analysis. This lesson discusses population, parameters, samples, and sampling methods of collecting data.

Population

In statistical language, the term *population* refers to the *total set of data* (actual or hypothetical) from which a sample is taken. Population, in statistical terms, applies to arbitrarily defined groups. It may refer to objects being measured or to measurements themselves. For example, a population may consist of all KC–135s in the Air Force inventory or a population could be all KC–135s at a certain base. Populations can be any size, from an unlimited number of items down to a dozen or so. Whether the population is real or imaginary, samples are taken from the source.

Parameters

In discussing the relationship between a population and a sample drawn from that population, you must be able to distinguish between characteristics of the population and corresponding characteristics of the samples. The mean, median, standard deviation and range of a population are characteristics of the population. Measures that characterize a population are called *parameters*.

Parameters describe the population. If you measure all individuals in a population, you can determine its parameters. Since it is usually impossible or impractical to measure a whole population, the parameters are usually unknown. Population parameters exist, however, whether they are known or not. You can take a sample from an unknown population, study the sample, and estimate population parameters.

Sample

A sample may be defined as part of a population or a part of the whole. There are two basic types of samples—random samples and biased samples. A *random* sample is a sample taken in such a manner that *each* value in the population has an *equal* chance of being selected. When certain individuals in a population have a better chance than others of being included in the sample, the sample is said to be *biased*. Most of the statistical procedures discussed later in this volume are based on random samples.

Random samples

A random sample is a sample where each item in a population has an equal chance of being included in the sample chosen. A random sample is one that has been picked "fairly" without the chances of any item within the population not being chosen. For example, suppose you're doing a study on maintenance man-hours expended on all F–16 aircraft, and you wish to take a sample of all the work orders (let's say about 50) that have been completed on this type of aircraft. You would not arrive at

a random sample by using your unit's Integrated Maintenance Data System (IMDS) database to select the work orders. Why not? First, when you use your unit's database, you are not giving every item in the population a chance of being included, since there are F–16s located at other bases. If you were to select items randomly from all the work orders recorded against F–16s, then you have a random sample.

Sampling techniques

The three types of selection methods in random sampling are stratified, cluster, and systematic.

Stratified sampling

Stratified sampling means dividing the population into subgroups (strata) in such a way that (1) there is as great a homogeneity (likeness) as possible within each stratum, and (2) there is as great a heterogeneity (difference) as possible between each stratum. After strata are set up, random samples are drawn that include data from each subgroup in proportion to its relative size. Stratified sampling is valuable when the distribution of the population is skewed (nonsymmetrical). Where such skewness exists, a small number of very important items may be included in the tail of the distribution that would likely be missed by unrestricted random sampling. Figure 2–1 shows the likeness of a stratified sample.

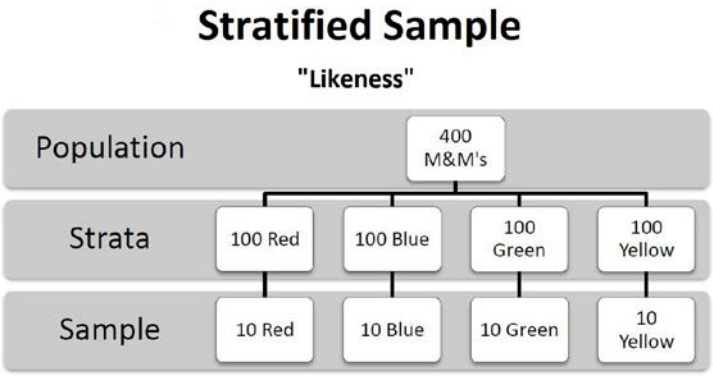


Figure 2–1. Stratified sample.

Cluster sampling

Cluster samples are also divided into subgroups but should display maximum heterogeneity (difference) within subgroup and maximum homogeneity (likeness) between subgroups. Cluster sampling has the *advantage* of ease of use when only portions of the population are available and when sampling by time periods. One important *disadvantage* of cluster sampling is that measures of variability from cluster samples are generally not as precise as when using other methods. The reason is that the heterogeneity of data within subgroups is rarely fulfilled since similar data tends to group together. Figure 2–2 shows the difference in a clustered sample.

Systematic sampling

Systematic sampling requires arranging the population in some systematic order and then selecting every Nth item until a given sample size is reached. One such example is constructing a query language processor (QLP) retrieval to select every 10th record.

The main *advantages* of systematic sampling are simplicity and saving time. However, systematic sampling may lead you astray if there are patterns in the population, such as a particular hour of the day that normally shows low productivity. An unlucky systematic sample that fits these hours is not representative of the whole population.

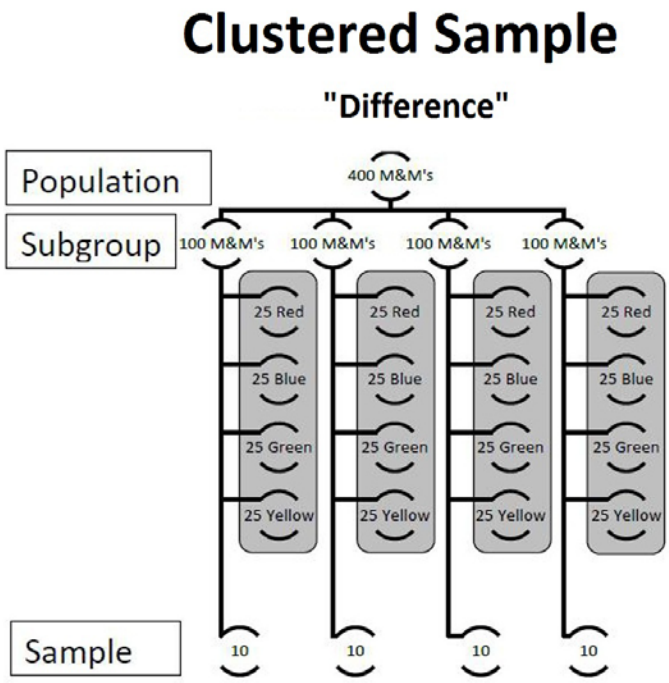


Figure 2–2. Clustered sample.

Bias samples

Bias samples are those that are chosen by some sort of selection, either knowingly or unknowingly. Remember the study involving the F–16s discussed earlier concerning random samples? Well, because you choose to use data only from your unit instead of from all units that have F–16s, this would be called *bias sampling*. Bias sampling can be further defined as unintentional bias or purposeful bias. *Unintentional bias* occurs when you accidentally distort a sample due to a lack of sample planning or forethought. Use *purposeful bias* sampling when you desire data for comparison studies that must meet certain qualifications, such as specific temperature ranges, skill levels, time periods, and so forth.

Statistics

Statistics are mathematical techniques or methods used in the collection, analysis, and interpretation of quantitative data. You use statistics to describe samples. Measures such as a mean, median, mode, and standard deviation are called *statistics* when they are computed from a sample. You are probably wondering "What are statistics good for?" Well, here are a few possibilities:

1. You can calculate averages to show a true picture of the typical performance of a group (population), such as the average cancellation rate for a fleet of KC–135s. The average calculated represents the entire KC–135 fleet.
2. You can determine the variability of the measurements. By using the average as a point of reference, you can determine how the cancellation rates for each KC–135 aircraft spread about this central point (average). You can then prepare graphs, tables, and figures to portray clearly the productivity of KC–135s.
3. The "raw" scores can be transformed into more meaningful form such as percentiles and standards.
4. You can determine the relationship of one variable to another. These statistics are called *correlation coefficients*. For example, you might relate variables such as personnel skill level with maintenance reliability. In other words, a statistic is a characteristic of a sample of data.

Statistics describe the sample, while the corresponding parameters describe the population. You will also be using Arabic letters ("a, b, c, d...") as symbols for statistics. Greek letters are used as symbols for parameters. For example, you represent the mean of a population with the Greek letter μ (*mu*), whereas you represent the mean of a sample taken from that population with $\overline{X}$. You represent the standard deviation of the population by the lower case Greek letter σ ("sigma"); however, you represent the standard deviation with a lowercase s.

406. Measurement scales

It is common to describe four types of data measures. The four types of measurement scales, in rank order from weakest to strongest, are *nominal*, *ordinal*, *interval*, and *ratio*. Refer to figure 2–3 as you complete this lesson.

Rank Order	Classification	Description
Nominal	Nonquantitative	Categories
Ordinal	Neither nonquantitative nor quantitative	Ranks
Interval	Quantitative	Arbitrary zero; equal intervals between values
Ratio	Quantitative	Absolute zero; equal intervals between values

Figure 2–3. Data measurement scales.

Nominal scale

Measurement at its weakest (or lowest) level exists when people, objects, or characteristics are placed in classes or categories that are different, with no higher than or lower than relationship between categories. The items placed in the categories can be counted, but they cannot be arranged in rank order. They are simply different. For example, people can be classified as male or female. The categories, male and female, make up a nominal measurement scale. Other examples of nominal scales are action taken codes (R, S, T), marital status (married, single, widowed, divorced, separated), branch of service (Army, Navy, Air Force, Marines, Coast Guard), shifts (A, B, C, D), type of employee (military, civilian), and so forth. Nominal scales are sometimes referred to as truly nonquantitative. Numbers or codes are often used to identify the various categories, but the numbers or codes have no quantitative significance. Nominal scales only indicate that items belong to different categories.

Ordinal scale

An ordinal scale is a somewhat stronger measuring device than a nominal scale. When a measurement scale consists of categories that are not only different but also stand in some sort of higher than or lower than relationship to each other, it's called an ordinal scale. The categories or classes in an ordinal scale are arranged in rank order, but the differences between the various ranks are not equal. For example, three mechanics may be ranked first, second, and third according to their abilities. The three ranks make up an ordinal scale. The scale tells you that the first mechanic is better than the second and that the second is better than the third. But such a scale cannot tell you if the difference between the first and the second is the same as the difference between the second and the third. Other examples of ordinal scales are school grades (A, B, C, D) and military ranks. Ordinal scales can indicate that objects are larger or smaller than another. An ordinal scale is neither truly nonquantitative nor truly quantitative.

Interval scale

When a measurement scale has *equal* intervals between scale values and an *arbitrary* zero point, it is an interval scale. The interval scale is the first truly quantitative scale that we have discussed. Because of the equal intervals between scale values, data from an interval scale may be added and subtracted. A good example of an interval scale is a Celsius thermometer. Since the zero point was

arbitrarily placed at the freezing point of water, the temperature can go below zero. You can say that a change in temperature from 0 to 20 is the same as the change from 20 to 40. But, because of the arbitrary zero point, you cannot say that 40 is twice as much temperature as 20. Another example of an interval scale is a Fahrenheit thermometer. It also has an arbitrary zero point, which is located at a point 32° Fahrenheit below freezing. A 0° reading does not represent an absence of temperature on either a Celsius or the Fahrenheit thermometers. The interval scale is stronger than the ordinal scale but not quite as strong as the ratio scale.

Ratio scale

A measurement scale that begins with a true or absolute zero point and also has equal intervals is a ratio scale. For example, weight is measured on a ratio scale. Because of the true zero point, you can say that an object weighing 4 pounds weighs twice as much as an object weighing 2 pounds. All measurements taken on this scale have a distinct relationship to the zero point, and so to each other. Measurements of *continuous data* made in feet, gallons, clock hours, and man-hours are also from *ratio scales*. The ratio scale, like the interval, is a truly quantitative measurement scale.

Data from some of the measurement scales that you have just read about should be subjected only to limited forms of statistical analyses. For example, finding the average of the categories of a nominal scale would result in a meaningless figure, because the number or codes are used only to name the categories. Averaging can only result in meaningful numbers when the data being averaged are from interval or ratio scales. Different statistical methods have different measurement requirements. Figure 2–3 depicts the major similarities and differences among the four measurement scales. Knowledge of the various measurement scales helps you understand the measurement requirements of the statistical techniques that you'll use later.

Self-Test Questions

After you complete these questions, you may check your answers at the end of the unit.

405. Data and samples

1. What do you call data that can be expressed only in whole number units?

2. What is the meaning of the term "population" as used in statistical language?

3. Define parameters.

4. What is a sample?

5. What is a random sample?

6. Define stratified sampling.

7. Explain the difference between purposeful and unintentional bias.

406. Measurement scales

1. Name two measurement scales that are truly quantitative.

2. Name one measurement scale that is truly nonquantitative.

3. Match each scale in column B with the most appropriate description in column A. Items in column B may be used only once.

Column A	*Column B*
____ (1) A measurement scale with equal intervals between scale values and an arbitrary zero point.	a. Nominal scale.
____ (2) A measurement scale consisting of categories that are different and cannot be arranged in rank order.	b. Ordinal scale.
____ (3) A measurement scale with scale categories that can be arranged in rank order, but the intervals between scale categories are not equal.	c. Interval scale.
____ (4) A measurement scale with equal intervals between scale values and a true zero point.	d. Ratio scale.

2–2. Data Distribution

A data distribution is simply a collection of data, usually about a particular item. Identifying the particular type of data distribution is extremely important in statistical analysis. For example, a statistical procedure that is appropriate for one kind of data distribution may not be appropriate for another.

407. Creating frequency distributions

When a collection of statistical data is gathered, it must be organized in a logical and usable fashion. This lesson covers frequency distribution characteristics and construction procedures. You will learn how to construct noncumulative and cumulative distributions. Using these distributions, you will easily be able to compute values and percentages to describe data distributions.

A frequency distribution is a tool used to summarize data and describe variation in performance. It is probably the most common way to organize data by combining the individual raw scores or values into a fewer number of categories and then summarizing the groupings into a *frequency table*. You can make numerous statistical calculations from frequency distributions when they have been constructed from data based on interval or ratio measurement scales.

Noncumulative frequency distribution

The frequency distribution data in figure 2–4 represents man-hours expended to remove and replace an avionics unit in B–1Bs at a certain base for the past year. The data, as shown here, is called raw data because it is not arranged in any particular fashion. Here are the procedures to construct a frequency distribution.

Step 1. Determine the range

The first step in making a frequency distribution is to *determine* the range of the data. The range, R, is the difference between the lowest value and the highest value. In this case, the lowest value is 2.1, and the highest value is 5.6. Therefore, the range is 3.5.

3.4	2.2	3.4	2.1 (low)	2.5
3.5	2.7	2.9	3.4	2.9
3.6	4.0	3.2	2.5	3.2
5.4	5.1	3.9	4.4	4.6
2.3	2.8	2.9	3.2	4.2
3.4	3.0	3.1	3.9	2.2
2.9	5.6 (high)	3.7	4.6	5.2
3.5	2.7	2.8	3.3	3.7
3.8	3.5	3.4	2.6	3.2
2.2	2.9	2.3	3.7	3.5

Figure 2–4. Man-hours required to replace B-1B avionics unit.

Step 2. Determine the class interval size

A class interval depicts the number of data subgroupings (classes) based on the established interval for a specified range of data. Since the recommended number of classes is from 10 to 20, pick a class interval size that's within this group. The class interval size can be determined by the formula:

$$\text{Interval} = \frac{\text{Range}}{\text{Desired number of classes}}$$

Based on a range of 3.5 (fig. 2–4), if you want to have 12 classes, which is within the 10 to 20 recommended number of classes, you will have an interval of 0.3. So in this case, you would use 0.3 for a class interval.

Consider the following additional points when selecting the class interval:

1. If the class interval chosen results in wide gaps between items falling in each class, the class interval is too large and the number of classes is too small. A basic assumption in the construction of frequency distributions is that the underlying pattern the data assumes in the mass is displayed by the distribution of the data classes.
2. If the class interval is too small, you lose smoothness and simplicity and are left with a ragged distribution. Large class intervals result in a *loss* of detail by lumping too much data into a single class.
3. Certain types of data may concentrate at certain values, which tends to group the data at specified intervals and forms natural class intervals. In addition, the midvalue, which represents the average value of the class, must coincide with these groupings or concentrations.
4. Do not allow class boundary values to overlap with adjacent class values. For example, a class should not be labeled 5 through 10 if the next class up is labeled 10 through 15. As you can see, the value 10 fits into either of the two classes, which makes it impossible to determine where to place an individual value of exactly 10.

Step 3. Set class limits

The third step in making a frequency distribution is to list the lower and upper limits of the class intervals (e.g., 2.1–2.3, 2.4–2.6, etc.), as shown in the "class" column of figure 2–3. It is customary to start with a bottom lower limit that is a multiple of the class interval (2.1–2.3). Also, the bottom class must contain the lowest value in the raw data (2.1). Note the lower limit of 2.1 for the bottom class in figure 2–3 that is a multiple of the 0.3 interval.

Step 4. Place tally marks

The last step in making a frequency distribution is to take the individual values (fig. 2–4), one at a time, and record them by placing tally marks to the right of the appropriate class, as shown in figure 2–5.

Class	Tally Marks	Frequency (f)
5.4-5.6	1 1	2
5.1-5.3	1 1	2
4.8-5.0		0
4.5-4.7	1 1	2
4.2-4.4	1 1	2
3.9-4.1	1 1 1	3
3.6-3.8	1 1 1 1 1	5
3.3-3.5	1 1 1 1 1 1 1 1 1 1	10
3.0-3.2	1 1 1 1 1 1	6
2.7–2.9	1 1 1 1 1 1 1 1 1	9
2.4–2.6	1 1 1	3
2.1–2.3	1 1 1 1 1 1	6
		N = 50

Figure 2–5. Frequency distribution of data.

Next, total the tallies for each class and place this number in the frequency column. Last, summarize the frequencies in column f to determine N, which is the total number of individual values in the distribution. To check your work, ensure that the total of N is equal to the number of items of raw data used. The midpoint value of any class is located halfway between its own lower limit and the lower limit of the class above it (e.g., the midpoint of class 2.4–2.6 is 2.5).

Cumulative frequency distribution

Up to this point, you have studied frequencies belonging to specific class intervals. Occasionally, you'll need information about the relationship between individual distribution values (or frequencies) and the total distribution value. In such situations, you may readily obtain this information by developing cumulative or percentage frequency distributions.

A cumulative frequency distribution displays the number of values falling above or below a certain point on a measurement scale (fig. 2–6, columns D and E). The cumulative frequency corresponding to the lower limit of any class interval is the number of individual values located above or below the lower limit of that class interval.

One way of comparing class frequencies to the total frequency is by percentage. You can determine percentage frequencies by dividing the individual class frequency by the total frequencies involved. Percentage frequencies readily show what part of the total each class represents (fig. 2–6, column C.)

A cumulative frequency distribution may also be combined with a percentage frequency distribution to form a percent of cumulative frequency distribution (fig. 2–6, columns F and G).

As you may have noticed, cumulative frequency distributions may be of the “less than” type (columns E, G) or the “more than” type (columns D and F). Since the distribution may be depicted two ways, always specify the type of distribution when constructing a frequency distribution table. Furthermore, frequency distribution tables need not always be developed to the extent of the one shown here, but only to the extent necessary to evaluate the data under study.

A	B	C	D	E	F	G
Class	F	% of Total	Cum. f Above	Cum. f Below	Cum. % Above	Cum % Below
5.7		0	0	50	0	100
5.4	2	4	2	48	4	96
5.1	2	4	4	46	8	92
4.8	0	0	4	46	8	92
4.5	2	4	6	44	12	88
4.2	2	4	8	42	16	84
3.9	3	6	11	39	22	78
3.6	5	10	16	34	32	68
3.3	10	20	26	24	52	48
3.0	6	12	32	18	64	36
2.7	9	18	41	9	82	18
2.4	3	6	44	6	88	12
2.1	6	12	50	0	100	0
	N = 50	100				

SI105358134

Figure 2–6. Cumulative frequency distribution.

408. Understanding graphic representations of frequency distributions

A graphic representation of a frequency distribution depicts data in a scaled model. From this model you can make generalizations about the data without lengthy computations. The arrangement of the tally marks in figure 2–7 gives you a general picture of the distribution of individual values. For example, the most frequent values fall in the class interval with a lower limit of 3.3, and the individual values are bunched in the lower half of the range. If you can imagine those tally marks on their side and reversed from left to right, you get a picture of a *histogram*. Another way to represent the data graphically is by use of a *frequency polygon*.

Histograms

Figure 2–7 shows a histogram depicting a graphical representation of the frequency distribution shown in figure 2–5. A histogram consists of a series of rectangles. Each rectangle represents one class of data. The lines that divide one rectangle from another are usually not projected down to the baseline. Notice that the histogram starts on a lower limit, gives a step-wide change from interval to interval, and ends on a lower limit. Histograms are based on the *assumption* that the individual values falling within each interval are *evenly* distributed over the interval. The total area of the histogram represents the total number of individual values in the distribution.

Frequency polygons

A frequency polygon representing the frequency distribution in figure 2–5 is shown in figure 2–8. Frequency polygons are *easier* to construct than histograms. When constructing a frequency polygon, you will plot the frequencies of the various class intervals *with* the corresponding midpoints. You will then use straight lines to connect the plotted points. To bring the ends of the frequency polygon down to the baseline, start the frequency polygon at the midpoint of the class just below the first one in which any individual value lies. End the frequency polygon at the midpoint of the class just above the last one in which any individual value falls.

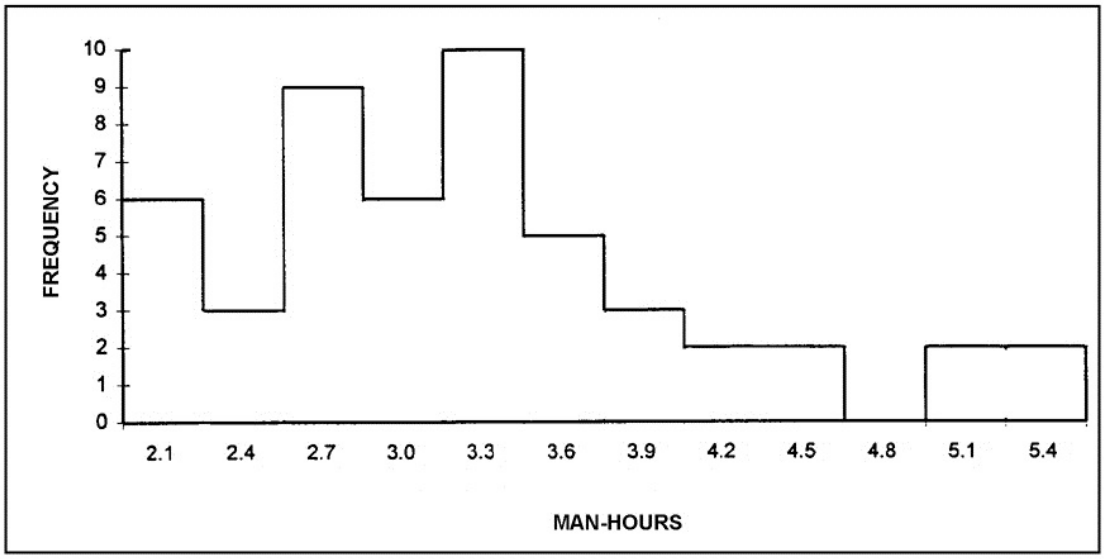


Figure 2–7. Histogram.

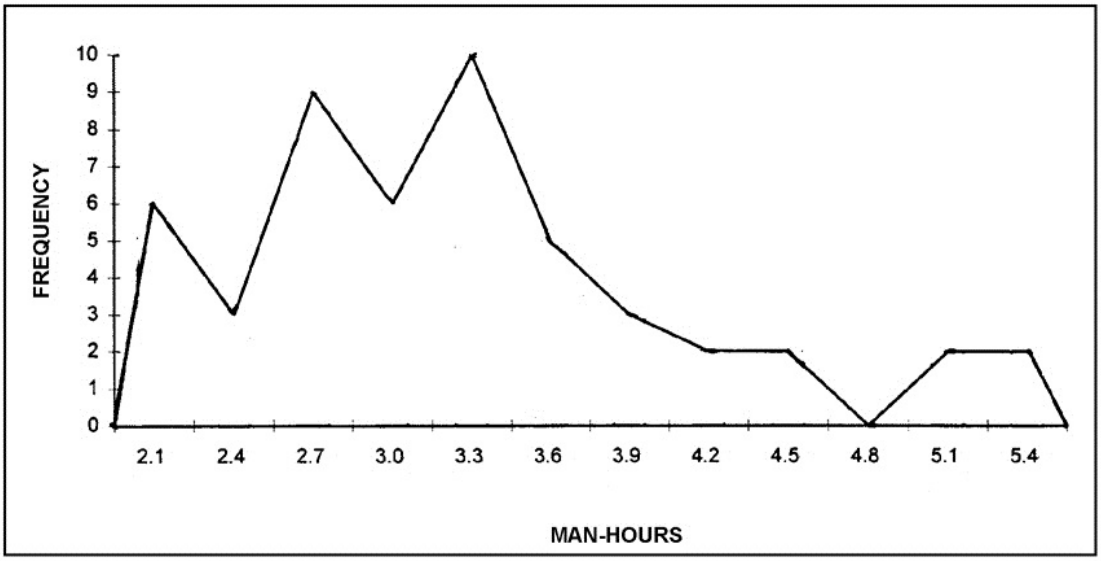


Figure 2–8. Frequency polygon.

Overlapping a histogram with a frequency polygon

A frequency polygon and a histogram representing the same distribution and drawn on the same axes are shown in figure 2–9. Notice that the histogram extends one-half class farther to the left and one-half class farther to the right than the frequency polygon. However, the areas under the two curves are the same. Notice also that for each section or area of the histogram cut off by the frequency polygon, an equal area or section is added.

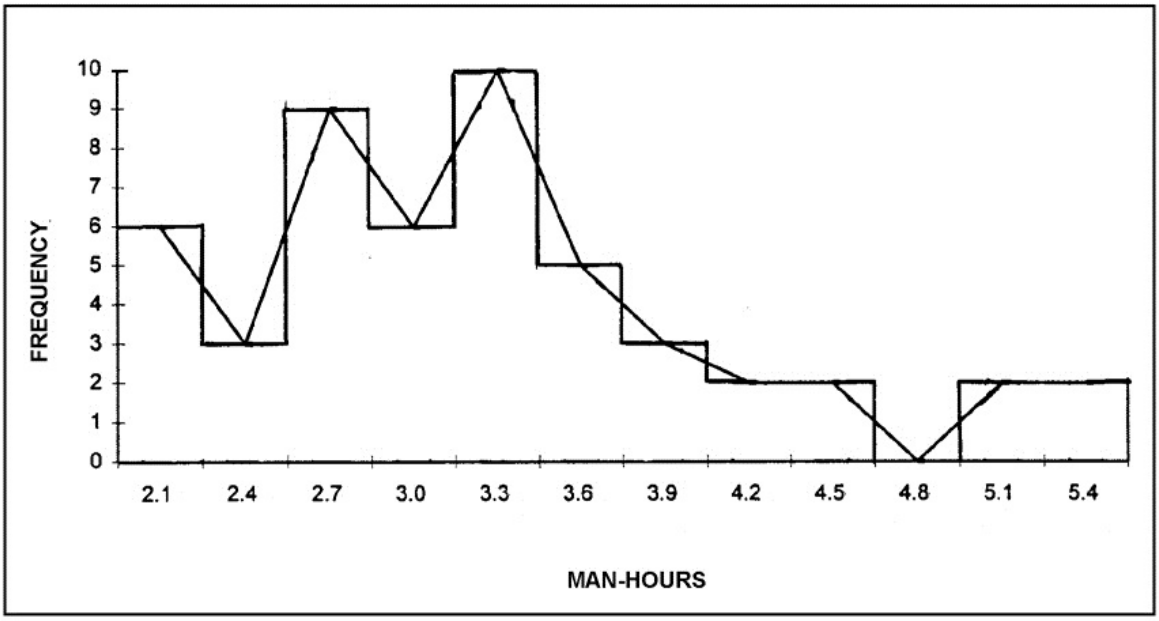


Figure 2–9. Histogram and frequency polygon.

The frequency polygon is easier to draw than the histogram and gives a *better* idea of the *general* shape of the distribution. On the other hand, the histogram gives a *better* representation of the number of *individuals* in each class. Histograms and polygons may be used not only with frequency of occurrence but also with other variables such as man-hours expended per time period. You'll find histograms and polygons to be of continuing use to you as an analyst.

Self-Test Questions

After you complete these questions, you may check your answers at the end of the unit.

407. Creating frequency distributions

1. What is the function of a frequency distribution?

2. What is the *first* step in making a frequency distribution?

3. What is the *recommended* number of classes in a frequency distribution?

4. What happens to a frequency distribution if the class interval is too small?

5. Explain cumulative frequency distribution.

6. How are percentage frequencies determined?

7. Refer to figure 2–6 to answer questions 7 through 10.

8. What percent of the equipment replacements required more than 4.5 man-hours?

9. What percent of the equipment replacements required between 3 and 3.9 man-hours?

10. Determine what percent of the equipment replacements required less than 3.6 man-hours.

11. How many equipment replacements took more than 5.1 hours?

408. Understanding graphic representations of frequency distributions

1. What are histograms based on?

2. Where are the frequencies of the class intervals plotted when constructing a frequency polygon?

3. Which graphic representation gives a better idea of the general shape of a distribution?

2–3. Measures of Central Tendency

One of the important ways of describing a group of measurements or scores is by the use of the measures of central tendency. The three measures of central tendency are: *mode*, *median*, and *mean*. In this section, you'll deal with the simplest application of these measures. If you recall from the last section, the majority of individual values in a frequency distribution cluster or "pile up" into a particular region. The location of this clustering is called the *central tendency of the distribution*. A measure of central tendency is a measurement that summarizes the individual values in the distribution and helps describe the data as a whole. Also, it enables you to compare two or more activities in terms of typical performance.

409. Determining the mode

The mode (*mo*) is the *most* frequently occurring value in a distribution. Since it occurs most frequently, it's the *most* usual or typical value. The mode is a good measure of central tendency to use *when* a rough estimate will do and a quick average is needed.

Characteristics

Here are five characteristics of the mode:

1. It is the *most* usual or typical value.
2. Its value is *not* affected by extreme values. The extreme values can be changed in size or eliminated without affecting the mode, *unless* the mode is one of those extreme values.

3. It is *simple* to estimate.
4. It is very *unstable*. In some cases, there are two or more modes. If there is no most frequently occurring value, the mode does *not* exist.
5. It is the *most* appropriate measure of central tendency to use with data from a nominal scale.

Finding the mode from ungrouped data

Of all measures of central tendency, the mode is the *easiest* to determine because it is obtained by observation—*not* computation. There are several methods for determining the mode. The method used will depend on the situation.

If you are trying to determine the mode from *ungrouped data*, all you have to do is find the value that occurs most frequently. For example, the mode of the values 3, 4, 4, 5, 5, 5, 6, and 9 is 5, because 5 is the *most* frequent value. Sometimes there is no most frequent value; in which case you *cannot* estimate the mode.

When the data has been grouped and placed on a *frequency distribution*, the class having the *highest* frequency is referred to as the *modal* class. The modal class is the class containing the mode. The midpoint of this class can be taken as a *quick* estimate of the mode. An example of the modal class of the frequency distribution in figure 2–10 is the class having a lower limit of 25. The midpoint of this class is 27.5. Thus, 27.5 is a quick estimate of the mode.

Suppose you are working with a frequency polygon or a histogram when you need an estimate of the mode. On a *frequency polygon* the mode is taken as a point on the baseline directly below the highest point on the frequency polygon (fig. 2–10). On a *histogram* the mode can be estimated by drawing dotted lines, also shown in figure 2–10, and drawing a line perpendicular to the baseline from the point where the dotted lines cross.

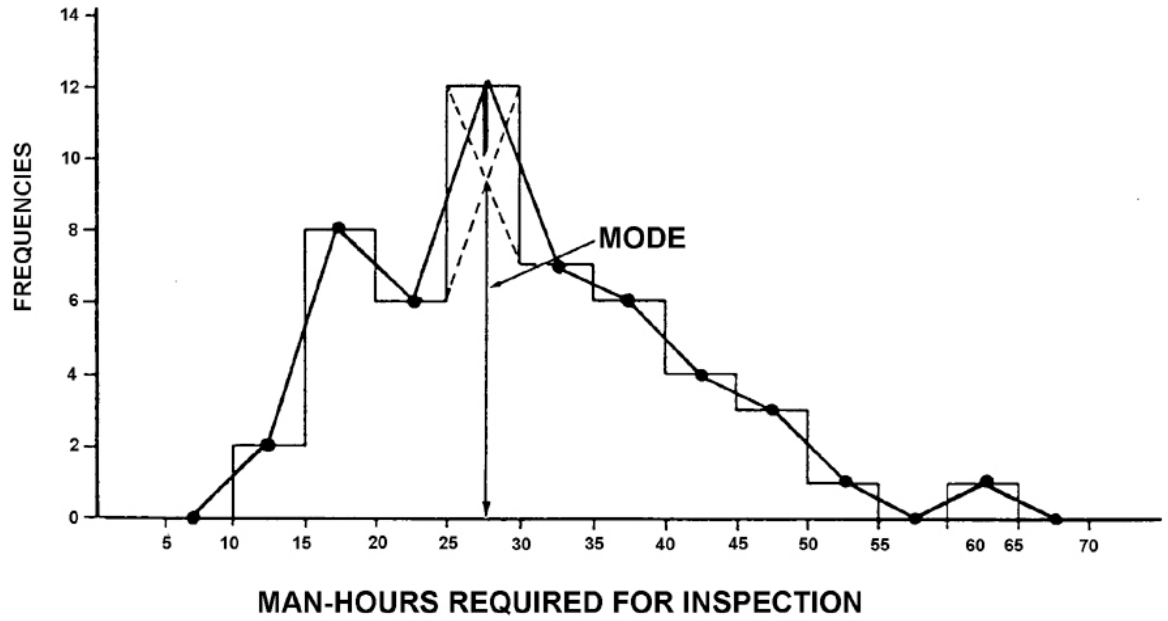


Figure 2–10. Frequency polygon and histogram showing the mode.

You have seen how you can quickly estimate the mode in various situations. There are more elaborate methods that can be used for estimating the mode, but the quick, easy methods give results that are *usually* satisfactory for analyzing maintenance data.

410. Determining the median

The median (*md*) is a point in the distribution that has 50 percent of the individual values on each side. Of the three measures of central tendency, the median is the *least* known. Even though the median is less common than the mean, analysts use it frequently because it's easy to compute and gives a better picture of the data than the mean and mode when data are *skewed*. For example, in a distribution of man-hours required to complete a unit of work, 50 percent of the jobs required fewer man-hours than the median and the other 50 percent required more. On a histogram or a frequency polygon, a vertical line drawn through the median splits the series in half.

Characteristics

Here are the six characteristics of the median:

1. It is an average of position on a measurement scale. As you have seen, it divides the distribution into two equal parts.
2. A value selected at random from a distribution is just as likely to be located above the median as below it.
3. Values *must* be arranged according to size *before* the median can be found.
4. It is *affected* by the total number of items, *not* by the size of extreme items. For example, the median of 2, 6, 8, 10, and 11 is 8. The extreme values may be changed so that the numbers are 5, 6, 8, 10, and 18 and the median is still 8.
5. It is not as familiar as the arithmetic mean to most people.
6. It can be used with data from ordinal, interval, or ratio measurement scales. It *cannot* be used with data from the *nominal* scale.

Finding the median from ungrouped data

To find the median of a series of ungrouped data, first arrange (array) the individual values in ascending order (lowest to highest). If there is an *odd number* of values, the middle value is then taken as the median. For example, you have a series of data as follows: 2, 8, 6, 5, 10, 13, and 4. Arrange the values in ascending order: 2, 4, 5, 6, 8, 10, 13. Since there is an odd number of values, the middle value, 6, is taken as the median. When there is an *even number* of values, the median is assumed to lie halfway between the two middle values. A series such as 3, 6, 8, 9, 10, 12 has an even number of values; therefore, the two middle values are 8 and 9. The median is assumed to be 8.5, which is halfway between the values 8 and 9.

411. Computing the mean

The mean (*m* or $\overline{X}$) or arithmetic average is the most widely used measure of central tendency. Basically, the mean is the sum of all the individual values divided by the total number of values. For example, the mean of the values 4, 6, 8, 10, 12 is

$$\overline{X} = \frac{4+6+8+10+12}{5} = 8$$

The four *most* commonly used formulas for computing the mean are the *arithmetic*, *weighted arithmetic*, *harmonic*, and *weighted harmonic*. Each method gives a slightly different answer, so you must carefully select the formula that is appropriate for each situation.

Arithmetic mean

The arithmetic mean is the most common measure of central tendency. It is the *arithmetic average* of a group of numbers. You use this when you want to place equal emphasis on each value in the data distribution.

Mathematical definition

A more precise definition of the arithmetic mean is *that point in a data distribution about which the sum of the deviations equals zero*. A deviation is a measure of how much a certain data point varies from a set point, in this case, the mean. For instance, assuming a group of data has a mean of 5, and one of the data points is 8, then the deviation of that data point from the mean would be +3. Refer to figure 2–11 for the following discussion.

(1)	(2)	(3)
X	$X - \overline{X}$	$(X - \overline{X})^2$
12	4	16
10	2	4
8	0	0
6	–2	4
4	–4	16
$\Sigma X = 40$	$\Sigma(X - \overline{X}) = 0$	$\Sigma(X - \overline{X})^2 = 40$
$\overline{X} = 8$		

Figure 2–11. Characteristics of the mean.

In a data distribution of five values (4, 6, 8, 10, 12), the mean is found to be 8. You determine this by summing the five values (40) and then dividing by the number of values (5). This is shown in column 1 of figure 2–11. X represents the data points. $\overline{X}$ is the mean, and the symbol Σ means "the sum of." Column 2 shows the deviations. Notice that the sum of these deviations is zero. A plus deviation indicates the data point is greater than the mean and a minus deviation indicates the data point is less than the mean. Column 3 is the sum of the squares of the deviations. Experimentation would show that if you picked any other data point in the distribution, calculated the deviations from that point, and summed the squares of those deviations; the sum would always be greater than that sum for the mean. For the data in figure 2–11, the sum of the squared deviations about any other point would be greater than 40. This then leads to another definition of the mean: *The mean is that point in a data distribution about which the sum of the squares of the deviations is at a minimum.*

Characteristics of the arithmetic mean

There are five true characteristics of the arithmetic mean:

1. The sum of the algebraic deviations from the arithmetic mean is zero. We take into account the plus and minus signs (column 2, fig. 2–11).
2. The sum of the squares of the deviations from the mean is less than about any other point. In column 3 of figure 2–11, each deviation has been squared, and the sum of the squares of the deviations is 40. For these X values, the sum of the squared deviations about any point other than the mean is greater than 40.
3. The value of the mean is determined by every item in the distribution. Notice in figure 2–11 that if any X value is changed, both the sum of the X values and the value of the mean changes. Every X value counts.
4. It is *greatly affected* by extreme values. Note what happens to the mean in figure 2–11 when 12 (column 1) is replaced with 27. The sum of 4, 6, 8, 10, and 27 is 55, so the mean becomes 11, which is larger than 4 of the X values. Because of the influence of the one extreme, 27, the mean pulled away from the other four X values.
5. It can be used with data from interval or ratio scales. It should not be used with data from nominal or ordinal scales.

Weighted arithmetic mean

Up to now, equal emphasis has been given to each item in a series. This equal emphasis may be misleading if individual items have different importance. Use the weighted arithmetic mean when you wish to emphasize the importance or value of each item.

Harmonic mean

The harmonic mean is defined as the value whose reciprocal is the arithmetic mean of the reciprocal of the values. It is another statistical technique that, when applied, will yield accurate information. The harmonic mean takes into consideration the differences between items when making overall assessments. The harmonic mean is used *primarily* for averaging rates, particularly rates of time. Since rates may be expressed in either of two forms, decimal or percentage, it is very easy to make the invalid assumption that each item measured shares an equal proportion of all time consumed. The harmonic mean prevents you from making this assumption.

Weighted harmonic mean

Like the harmonic mean, the weighted harmonic mean provides an average of rate for a period of time, while at the same time providing weighted emphasis to each value.

Finding the mean from ungrouped data

It is easier to work with ungrouped data than to go through the laborious processes of grouping.

Arithmetic mean

For ungrouped data, the arithmetic mean is determined by simply finding the sum of the given values and dividing this by the number of values. The application of this principle can be expressed as $\overline{X} = \frac{\Sigma X}{N}$ where $\overline{X}$ is the mean, Σ is the symbol for "sum of" or summation, and N is the number of individual values or cases symbolized by X. Taking the values 1, 6, 8, 9, 13, 26, 28, you find ΣX = 91 and N = 7. Therefore, $\overline{X}$ = 91 ÷ 7, or 13. You can see that this operation is the same as the one that is frequently used in everyday life in figuring bowling averages, gas mileage, and a multitude of other averages.

Weighted arithmetic mean

If a certain worker can perform job A in 5 minutes, job B in 10 minutes, and job C in 20 minutes, what is the mean time for the jobs? If *only* one of each job is performed, the following formula applies:

$$\overline{X} = \Sigma\frac{X}{N} = \frac{5+10+20}{1+1+1} = \frac{35}{3} = 11.67 \text{ minutes}$$

However, if the worker does 4 of the A tasks, 3 of the B tasks, and 6 of the C tasks, then our series of Xs in the formula represent different sized elements. The total N of all elements now equals 13 instead of 3. To represent every job properly, the formula *must* read as follows:

$$\overline{X} = \frac{(4A + 3B + 6C)}{(4+3+6)}$$

$$\overline{X} = \frac{(4\times 5\text{ min}) + (3\times 10\text{ min}) + (6\times 20\text{ min})}{4+3+6}$$

$$\overline{X} = \frac{170\text{ min}}{13} = 13.08 \text{ minutes}$$

Notice that the numbers 4, 3, and 6 represent both the quantities of the jobs performed or "N" and also the "weights" of each job performed by type. This modifies the basic formula somewhat from:

$$\overline{X} = \frac{\Sigma X}{N}$$

to:

$$\overline{X} = \frac{\Sigma \text{ weighted } X}{\Sigma \text{ weights}} \text{ or } \overline{X} = \frac{\Sigma WX}{\Sigma W}$$

The weighted arithmetic mean is also *particularly* useful when you desire a *mean of means*. There are several important points to remember about weighted means. First, weighing places the correct emphasis on each item in a series according to its relative importance to the series. To apply the weight, multiply an item's value in the series by its appropriate quantity factor. Also, the arithmetic mean for a series cannot be accurately computed unless each item in the series is equally represented. Finally, to average two or more series' means together, the means of the individual series cannot have the same weight unless the items from which the series were derived are equal to each other in number. Therefore, when combining items of different data masses, do not average the means of the masses together; instead, treat the sum of all the individual values as one data mass and then take its average.

Harmonic mean

Many situations require you to use the harmonic mean instead of the arithmetic mean. As an example, suppose that three individuals complete a like task with the following results:

Airman A takes 30 minutes.
Airman B takes 20 minutes.
Airman C takes 40 minutes.

To determine the average time needed to complete one task, you can say 30 + 20 + 40 = 90 minutes divided by 3 people, for an average of 30 minutes per job. However, you can also say that:

Airman A completes 2 jobs per hour.
Airman B completes 3 jobs per hour.
Airman C completes 1.5 jobs per hour.

With the total time of 180 minutes divided by the total of 6.5 jobs, you have an average of 27.8 minutes per job. Which method is correct? Since the first method made the invalid assumption that all personnel complete an equal quantity of units, then obviously the second method is the correct way. By using the harmonic mean, you can *eliminate* the possibility of an *invalid* assumption. The formula follows:

$$H = \frac{N}{\Sigma\left(\frac{1}{X1} + \frac{1}{X2} + \frac{1}{X3} \ldots\right)}$$

In this case, it is:

$$H = \frac{3}{\frac{1}{30}+\frac{1}{20}+\frac{1}{40}} = \frac{3}{.033+.050+.025}$$

or:

$$H = \frac{3}{.108} = 27.8$$

As another example, assume that a person travels from point A to point B at 30 miles per hour and returns from B to A at 60 miles per hour. To find the average speed for the round trip, if the distance is assumed to be 60 miles, you can say:

$$\text{Time from A to B} = \frac{60 \text{ miles}}{30 \text{ miles per hour}} = 2 \text{ hours}$$

$$\text{Time from B to A} = \frac{60 \text{ miles}}{60 \text{ miles per hour}} = 1 \text{ hour}$$

$$\text{Average speed for the trip} = \frac{\text{total distance}}{\text{total time}} = \frac{120 \text{ miles}}{3 \text{ hours}} = 40 \text{ miles per hour}$$

The above average of 40 miles per hour is equal to the harmonic mean of 30 and 60, that is:

$$\frac{2}{\frac{1}{30}+\frac{1}{60}} = 40 \text{ miles per hour}$$

However, one might be tempted to take the arithmetic mean of 30 and 60 miles per hour to obtain 45 miles per hour, but this is incorrect.

As you can see, the harmonic mean can be used for various tasks in maintenance analysis. You'll find this method very useful in forecasting personnel capabilities, which are based on past data. But, if zeros are present in the series, the harmonic mean is useless.

Weighted harmonic mean

Like the arithmetic mean, the harmonic mean may be weighted when various quantities of the denominator factors are used. To illustrate this, we'll use the time rates in example 1 and increase the number of individuals performing, that is:

$$H = \frac{9}{3(\frac{1}{20})+4(\frac{1}{30})+2(\frac{1}{40})} = 27.1 \text{ minutes}$$

The harmonic mean is *always less* than the arithmetic mean.

Self-Test Questions

After you complete these questions, you may check your answers at the end of the unit.

409. Determining the mode

1. Define the mode.

2. List the five characteristics of the mode.

3. How is the mode estimated from ungrouped data?

4. Estimate the mode for the values 5, 6, 6, 8, 8, 8, 8, 8, 10, 12.

5. How can you estimate the mode from a frequency distribution?

6. How can you estimate the mode from a histogram?

410. Determining the median

1. Define the median.

2. List six characteristics of the median.

3. How is the median calculated from ungrouped data when there is an odd number of values?

4. Determine the median of 6, 3, 8, 4, 10.

5. How is the median calculated from ungrouped data when there is an even number of values?

6. Determine the median of 12, 3, 5, 6, 7, 15.

411. Computing the mean

1. When should you use the arithmetic mean?

2. What is the mathematical definition of the arithmetic mean?

3. List five characteristics of the mean.

4. Find the arithmetic mean for the values 8, 12, 4, 9, 7, 2.

5. What is the harmonic mean *primarily* used for?

6. How is weight calculated when using the weighted arithmetic means?

7. When would a harmonic mean be considered useless?

2–4. Measures of Variability

Previously, you learned the various measures of central tendency. These measures or averages do not by themselves adequately describe a data distribution. They locate the center of a distribution, but they show nothing about how the data is arranged around it. Measures of variability indicate whether the data items are close together or spread far apart—they measure the scatter or spread of the data. You'll learn about the standard deviation, standard error of the means, and normal and nonsymmetrical distribution curves.

412. Computing the standard deviation

The degree to which the individual values tend to spread about an average is called the *variability* of the data. It is a measure that expresses how much data in a distribution differ or vary from one another. Variability may also be called *dispersion, scatter, spread,* or *variation*. The measures of variability discussed here are standard deviation and variance.

One of the most useful measures of dispersion or variability is the *standard deviation*. It is the most widely encountered because it is used in so many statistical computations. There are two methods to compute the standard deviation, in this set of instructions the focus will be on the deviation method. The standard deviation is defined as *the square root of the squares of all the deviations from the mean*. It is the actual deviation or distance between every individual value from the mean. The value of standard deviation denotes how spread out the numbers are in a set of data.

You can mathematically define the standard deviation as the square root of the average of the squares of all the deviations from the mean. You can express this statement symbolically as:

$$S = \sqrt{\frac{\sum\left(X - \overline{X}\right)^2}{N - 1}}$$

Where:

S = standard deviation

X = specific/individual value

$\overline{X}$ = mean

N = number of values

Variance

The variance (s^2), which is closely related to the standard deviation, is simply the standard deviation squared. Although this formula is fairly easy to understand, it is not a form that is easy to use. Its use involves finding the mean, computing the deviation of each X value from the mean, squaring each of these deviations, averaging the squares, and then finding the square root (fig. 2–12). Notice that the standard deviation, 3.16, is affected by every X value in the distribution. If you change any of the X values so as to cause an increase in the variability, the standard deviation also increases because it is a measure of that variability. Of all the measures of variability, standard deviation is by far the most widely used. We use the formula in figure 2–12 only to help in explaining the meaning of the concept.

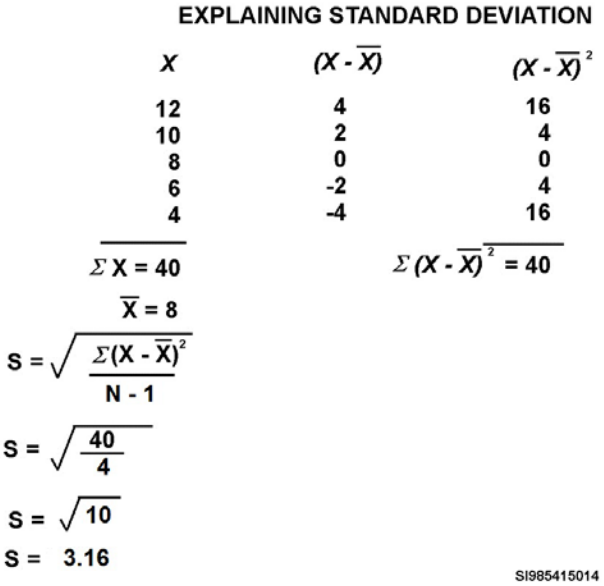


Figure 2–12. Explaining standard deviation.

We'll use an example of mission capable rates for the remainder of this lesson. Refer to figure 2–13 for the data items.

MC Rates, Airframe A X	Mean $\bar{X}$	Deviations $X - \bar{X}$
81	79	+2
80	79	+1
78	79	–1
77	79	–2
		$\Sigma(X - \bar{X}) = 0$
MC Rates, Airframe B X	Mean $\bar{X}$	Deviations $X - \bar{X}$
91	86	+5
86	86	0
85	86	–1
82	86	–4
		$\Sigma(X - \bar{X}) = 0$

Figure 2–13. Mission capable rates.

We used mission capable (MC) rates for two airframes. The figure displays the MC values, the means, and the differences between each value and the means. As you can see, the sum of the deviations for airframe A and airframe B is zero. You may also notice that the deviations for airframe A are less scattered than those of airframe B. If you calculate the means of the deviations, you will find that they are also zero: (–2 –1 +1 +2)/4 = 0/4 = 0 and (–4 –1 +0 +5)/4 = 0/4 = 0.

For any distribution, the *sum* of the deviations is *zero* and, therefore, the *mean* of the deviations is *zero*. Since we want to *compare* the data spread of data distributions, the means of the deviations would *not* be an adequate comparison indicator. Statisticians developed a procedure that uses the deviation measures yet avoids this "zero" problem. They first squared the deviations; this gets rid of the negative values. Then they obtained the mean of these values. Refer to figure 2–14, which uses the deviations from figure 2–13.

Deviations, Airframe A $X - \bar{X}$	**Squared Deviations** $(X - \bar{X})^2$
+2	4
+1	1
–1	1
–2	4
$\Sigma(X - \bar{X}) = 0$	$\Sigma(X - \bar{X})^2 = 10$
Deviations, Airframe B $X - \bar{X}$	**Squared Deviations** $(X - \bar{X})^2$
+5	25
0	0
–1	1
–4	16
$\Sigma(X - \bar{X}) = 0$	$\Sigma(X - \bar{X})^2 = 42$

Figure 2–14. Calculating standard deviations.

Computing the variance and standard deviation

The mean of the squared deviation is known as variance, represented by s^2. The formula for obtaining the variance is:

$$s^2 = \frac{\Sigma (\chi - \overline{X})^2}{N-1}$$

Where:

s^2 = variance

N = number of values

$\Sigma(X - \overline{X})^2$ = the sum of the squared deviations

The mean of the squared deviations of airframe A is 10/3, which 3.33, and the mean of the squared deviations of airframe B is 42/3, which equals 14. The variance is used in several statistical situations, but in this case, you still need to obtain a number that represents the original deviations and not the squares of those deviations. Quite simply, then, you take the square root of the variance ($\sqrt{s^2} = s$), which is called the *standard deviation*. The standard deviation of airframe A is the square root of 3.33, which is 1.82, and the standard deviation of airframe B is the square root of 14, which is 3.74.

Standard deviation formulas

Standard deviation can be calculated in two different ways, depending on whether you are calculating it from a population or a sample. Recall that a population is an entire set of data, and a sample is a selected subset of a population. We explain the difference as we go over the formulas.

Standard deviation of a population

When you have an entire population of data to analyze, use the formula for finding the standard deviation of a population. The following formula is used on ungrouped data for a population of values:

$$\sigma = \sqrt{\frac{\sum (X - \mu)^2}{N}}$$

Where:

σ = Standard deviation of a population

X = specific value

μ = population mean

N = the number of values

Say you have the values 3, 5, 6, 7, 9 as a population of ungrouped data. Applying the formula would give the following results:

$\Sigma X = 30$

$\Sigma(X - \mu)^2 = 20$

$N=5$

$$\sigma = \sqrt{\frac{\sum (X - \mu)^2}{N}}$$

$$\sigma = \sqrt{\frac{20}{5}}$$

$$\sigma = \sqrt{4}$$

$$\sigma = 2$$

The number 2 represents the measure of variability or scatter among the values of this population. By looking at the formula, you can see that a change in one of the X values changes the standard deviation. The formula also shows a *direct* relationship between the X values and the standard deviation; that is, as the X values *increase*, the standard deviation *increases*. Conversely, as the X values decrease, the standard deviation decreases. If *all* the values are the *same*, there is no variability and so the standard deviation is *zero*.

Standard deviation of a sample

Usually, you will find yourself dealing with samples rather than whole populations of data. The following formula calculates the standard deviation of a sample of ungrouped data:

$$s = \sqrt{\frac{\sum (X - \bar{X})^2}{N - 1}}$$

Where:

s = standard deviation

X = specific value

$\bar{X}$ = sample mean

N = number of values

Now let's take a sample from a population of data using the same values as before (3, 5, 6, 7, 9). Apply the following formula:

$$\Sigma X = 30$$
$$\Sigma (X - \bar{X})^2 = 20$$
$$N = 5$$

$$s = \sqrt{\frac{\sum (X - \bar{X})^2}{N - 1}}$$

$$s = \sqrt{\frac{20}{5-1}}$$

$$s = \sqrt{5}$$

$$s = 2.24$$

The number 2.24 is the standard deviation for our sample of values. This is slightly higher than the standard deviation of the same values representing a population. Why use a different formula? You must compensate for the fact that there is usually less variability in a sample as compared to the whole population. Using N–1 results in a smaller divisor that gives a larger result. In a normal distribution, as n grows *smaller*, s becomes *less* representative of the population (σ).

413. Computing the standard error of the mean

So far you learned how to measure the variability or spread of a frequency distribution made up of individual X values from one large sample of data. Imagine a frequency distribution made up of means ($\overline{X}$) from many small random samples and consider estimating its variability.

Sampling distribution of the mean

If a large number of random samples of a given size are taken from a population, and the mean of each sample is computed, the means ($\overline{X}$ values) of the samples will themselves form a frequency distribution. This theoretical frequency distribution of $\overline{X}$ values is called a sampling distribution of the means. Like all frequency distributions, this sampling distribution of the means has its own central tendency and variability.

Sampling theory gives some very definite information about the average and spread of the $\overline{X}$ values in a sampling distribution of means. The $\overline{X}$ values tend to be concentrated around the population mean (μ), and in the long run, the mean of the $\overline{X}$ values is *the same* as the population mean (μ). The mean of the means (or the average of the averages, as it is sometimes called) is symbolized by $\overline{\overline{X}}$. The variability of the $\overline{X}$ values in the sampling distribution of means is affected by both the variability of the population from which the random samples were drawn and the random sample size.

Standard error of the mean

The standard deviation of a sampling distribution of means is called the standard error of the mean and is equal to the standard deviation of the population divided by the square root of the sample size, that is,

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

Where:

σ = standard deviation of the population

n = sample size

$\sigma_{\overline{X}}$ = standard error of the mean

Therefore, the variability of the $\overline{X}$ values varies directly as the variability of the population and *inversely* with the sample size. For example, if the sample size is 4, the variability of the $\overline{X}$ values is only one-half as great as the variability of the population. If the sample size changes to 16, the variability of the $\overline{X}$ values is only one-fourth as great as the variability of the population.

There is a tendency for a distribution of sample averages to be normal, regardless of the form of the population from which the sample was drawn. If the population is a normal distribution, then the sample $\overline{X}$ values form a normal distribution in the long run. If the population is skewed, the distribution of $\overline{X}$ values tends to be less skewed than the population. This tendency of a sampling distribution of means to form a normal distribution—even though the samples are taken from a skewed distribution—is an advantage of the control charts for averages, which you'll study later.

Estimating standard error of the mean

In the above formula, the standard deviation of the population was given. However, in many practical situations in analysis, some parameters of the population are not known. When the population standard deviation or mean is not known, we can use an estimation of the standard error of the mean. Below we will discuss two methods of estimating the standard error of the mean.

Infinite population

When the population of the data is infinitely large, as in the millions, the standard deviation of the entire population σ is not known. Therefore, the best you can do is use the sample standard deviation, s, as an estimate of the population standard deviation.

There is a tendency for the standard deviation of a small random sample to be less than the standard deviation of the population from which the sample was taken. To correct for this tendency, n–1 instead of n is used in the formula for estimating the standard error of the mean when the sample size is small. As the sample size increases, the minus 1 correction factor has less and less effect on the answer. Therefore, you can forget about the minus 1 when the sample size is large (when n is 30 or more). An estimate of the standard error of the mean is symbolized as $s_{\overline{X}}$. The *two* versions of the formula for *estimating* the standard error of the mean are as follows:

1. For sample size of *less* than 30:

$$s_{\overline{X}} = \frac{s}{\sqrt{n-1}}$$

2. For sample size of 30 or *more*:

$$s_{\overline{X}} = \frac{s}{\sqrt{n}}$$

Consider the values 2, 3, 3, 4, 4, 4, 5, 5, 6, 7 as a small random sample and estimate the standard error of the mean. First, compute the standard deviation of the sample using the formula for ungrouped data as follows:

$$S = \sqrt{\frac{\sum\left(X - \overline{X}\right)^2}{N-1}}$$

$$s = \sqrt{\frac{20}{10-1}}$$

$$s = \sqrt{2.23}$$

$$s = 1.49$$

Then, use the standard deviation of the sample as an estimate of the population standard deviation and solve for an estimate of the standard error of the mean as follows:

$$n < 30$$

$$S_{\bar{X}} = \frac{s}{\sqrt{n-1}}$$

$$S_{\bar{X}} = \frac{1.49}{\sqrt{10-1}}$$

$$S_{\bar{X}} = \frac{1.49}{3}$$

$$S_{\bar{X}} = 0.49$$

As another example, treat 50 X values as a random sample taken from an unknown population of man-hours required for inspection and estimate the standard error of the mean. First, compute the standard deviation. Assume that this has already been done with a standard deviation of 10.75 for this sample. The standard error of the mean for the sample size of 50 is estimated as follows:

$$n \geq 30$$

$$\sigma_{\bar{X}} = \frac{s}{\sqrt{n}}$$

$$\sigma_{\bar{X}} = \frac{10.75}{\sqrt{50}}$$

$$\sigma_{\bar{X}} = \frac{10.75}{7.07}$$

$$\sigma_{\bar{X}} = 1.52$$

If the sample size is changed to a smaller number, then the standard error of the mean becomes larger and more fluctuation is expected in the distribution of sample means.

Finite population

When you pick a sample size of n from a large population size of N (but less than a million), the standard error of the mean can be calculated using a variation of the basic formula. This applies when you use a random sample size. An advantage of this formula is that you are not restricted by the less than 30 or more than 30-sample size rule. You multiply a factor, called the *finite population correction factor*, to the standard error of the mean for a population. The factor is as follows:

$$\sqrt{\frac{N-n}{N-1}}$$

Where:

N = size of the population

n = sample size (chosen at random)

You multiply this factor to our standard deviation of the means, thus, you get:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

This formula applies commonly when you pick a random sample from the population and use that sample for calculating the standard error of the mean. Let's say that you have a population of 1,000 data items. The standard deviation of the population σ equals 3. The sample size is randomly picked at 36. Use the following formula:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{3}{\sqrt{36}}\sqrt{\frac{1000-36}{1000-1}} = 0.5(0.982) = 0.491$$

For all practical purposes, when N becomes very large compared to the sample size, the *finite population correction factor* gets closer to 1. The standard error of the mean formula applied to a finite population depends on the population σ and the sample size n. The standard error of the mean will be large if the population is large and the sample size is small. The standard error decreases if you increase the sample size. Therefore, to reduce the standard error of the mean, you must increase the sample size—the larger the better. This will give us a better picture of the population.

When using standard error of the mean, remember that this measure does not describe how individual values are dispersed about a population mean, but rather how means of samples vary around the population mean. Stated another way, it provides a measure of the amount of error between a sample mean and the true population mean.

414. Define normal distribution curves

The normal distribution is one of the oldest statistical measurements. It was developed initially in the 18th century by scientists dealing with errors in measurement.

Describing a normal distribution

The distribution forms a symmetrical, bell-shaped curve. Symmetry dictates that the left and right halves of the curve are alike. Using the strength of individuals as an example, you can see how people fit this curve. There is a small number of very weak people and a small number of very strong people, but the majority of people in the middle have average amounts of strength. Though the tails of the distribution actually extend to infinity on each side, nearly all the applications of the curve limit the tails to certain useful values. Later in this lesson, you'll see the useful portion of the curve and the determination of these limits.

Parameters

A normal distribution has *two* parameters—the mean and the standard deviation. The *mean* is in the middle of the normal curve and tells the average value of the X variables. This measure is symbolized by $\bar{X}$ for a sample and μ for a population. The standard deviation, symbolized by s for samples and σ for populations, is a measure of the degree to which the data is spread out or dispersed from the mean. A large standard deviation indicates data that are spread over a wide range, whereas a small standard deviation shows data grouped closely about the mean. Over 99 percent of the total area under the curve is contained within *three* s values to the left and to the right of the mean. Because of this characteristic, we frequently neglect the extreme tails of the curve and concern ourselves only with the useful range. Figure 2–15 shows the normal distribution subdivided into segments by the standard deviation and points out the useful area. Of course, like all continuous distributions, the *total* area under the *curve* is 1 or 100 percent. With the two parameters, $\bar{X}$ and s, you have totally described the data distribution.

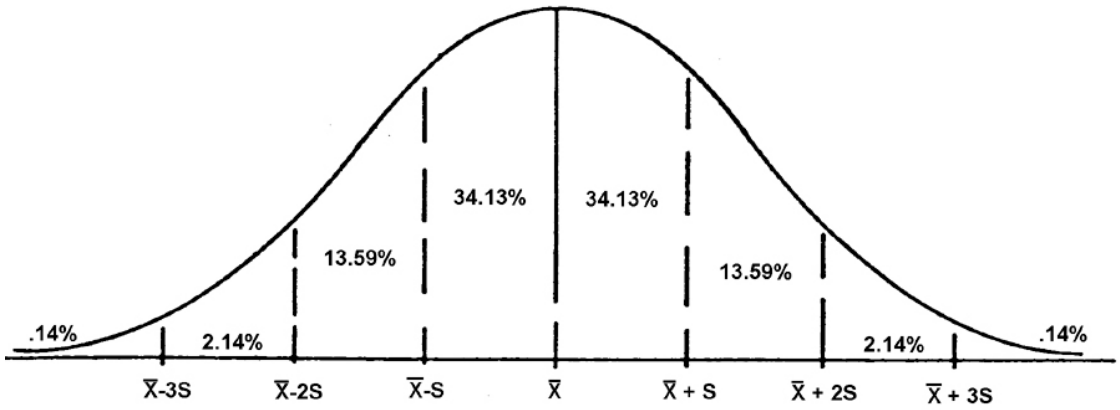


Figure 2–15. Normal curve showing standard deviations.

Sections

You can subdivide the area under the normal curve into *six* useful sections. Let's look at a maintenance example and use it to illustrate probability. Suppose the past data have indicated that a removal and replacement task follows a normal distribution with a mean time of 3.0 hours and a standard deviation of 0.5 hours. Figure 2–16 shows a normal curve with six subdivisions on the x scale. In this example, x is a time variable. The percentage values in the curve are the percent of the total area that is contained in each subdivision. Note that the total area is 100 percent.

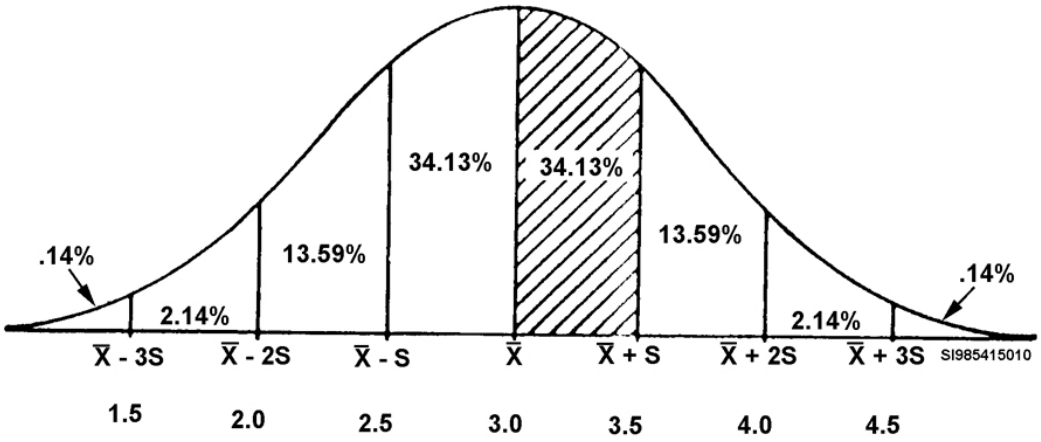


Figure 2–16. Normal curve with subdivisions.

Figure 2–17 contains a table that lists normal curve probabilities between the mean ($\overline{X}$) and any given z score value. This table is called a "normal curve tail area table;" it is the most commonly used table in analysis. This table lists probabilities beyond the z score value away from X, or stated differently, the probabilities in one tail of the normal curve. You use the normal curve tail area table to answer probability questions of "greater than" or "less than" for any given X value in the distribution. The z value for 0.0 in this table is .5000. This is a clue that you are using the proper table.

Areas In One Tail of the
Normal Curve at Selected values of z

Z	0.0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	**.0418**	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010

NORMAL CURVE TAIL AREA TABLE (Right Tail)

Figure 2–17. Normal curve tail area table.

Normality

Now you will study the concept of normality, because many statistical methods are based on the assumption that the data mass is normally distributed around the mean. The normal distribution is extremely useful to maintenance data systems analysis. Its characteristics and probability solutions are actually quite simple and easy to describe and solve. To use it, however, you must first determine whether the data really is from a normal distribution. There are several ways of doing this, but the easiest by far is the graphic technique.

The concept behind the graphic technique is that data from a normal distribution shows up as a *straight line* when plotted on normal probability graph paper. The normal probability paper, as shown in figure 2–18, may be used. Notice that the vertical scale is a straight arithmetic scale, while the horizontal scale is an actual normal distribution. The range of the horizontal scale is from 0.01 to 99.99 percent. This is well within the range of the useful part of the normal curve, which you studied previously. Study figures 2–18 and 2–19 as you go through the steps of plotting the data.

Step 1. Arrange the data

The first step in plotting data is to arrange the data in an array from the lowest to the highest value. Next, construct a value scale along the arithmetic side of the graph paper large enough to contain all of the values in the distribution. Number the individual values consecutively from 1 to n, the sample size, and convert each number to its respective percentage value by dividing the number by n + 1. The reason n + 1 is used instead of n when computing percentages is to prevent overextending the graph paper. This example has only 9 values in order to simplify the process, but any number of individual values may be used. You are now ready to plot the data.

Step 2. Plot the data

Beginning with the lowest value, find its position on the arithmetic scale and then read across horizontally until you come to its respective percent value, where you make the plot. Continue in this manner with each value in the sample until all of the values have been plotted.

Step 3. Analyze the data

Now you are ready to make a decision regarding the normality of the data. If the points fit an approximate straight line, or if a straight line fits the points, then you can reasonably assume that the data is normally distributed. Figure 2–19 shows the data does indeed meet this criterion.

Raw Data	Array	Rank	% of n + 1
11	6	1	10
22	11	2	20
6	14	3	30
14	19	4	40
36	22	5	50
40	23	6	60
23	30	7	70
19	36	8	80
30	40	9	90
		n = 9 n + 1 = 10	

Figure 2–18. Arrangement of data.

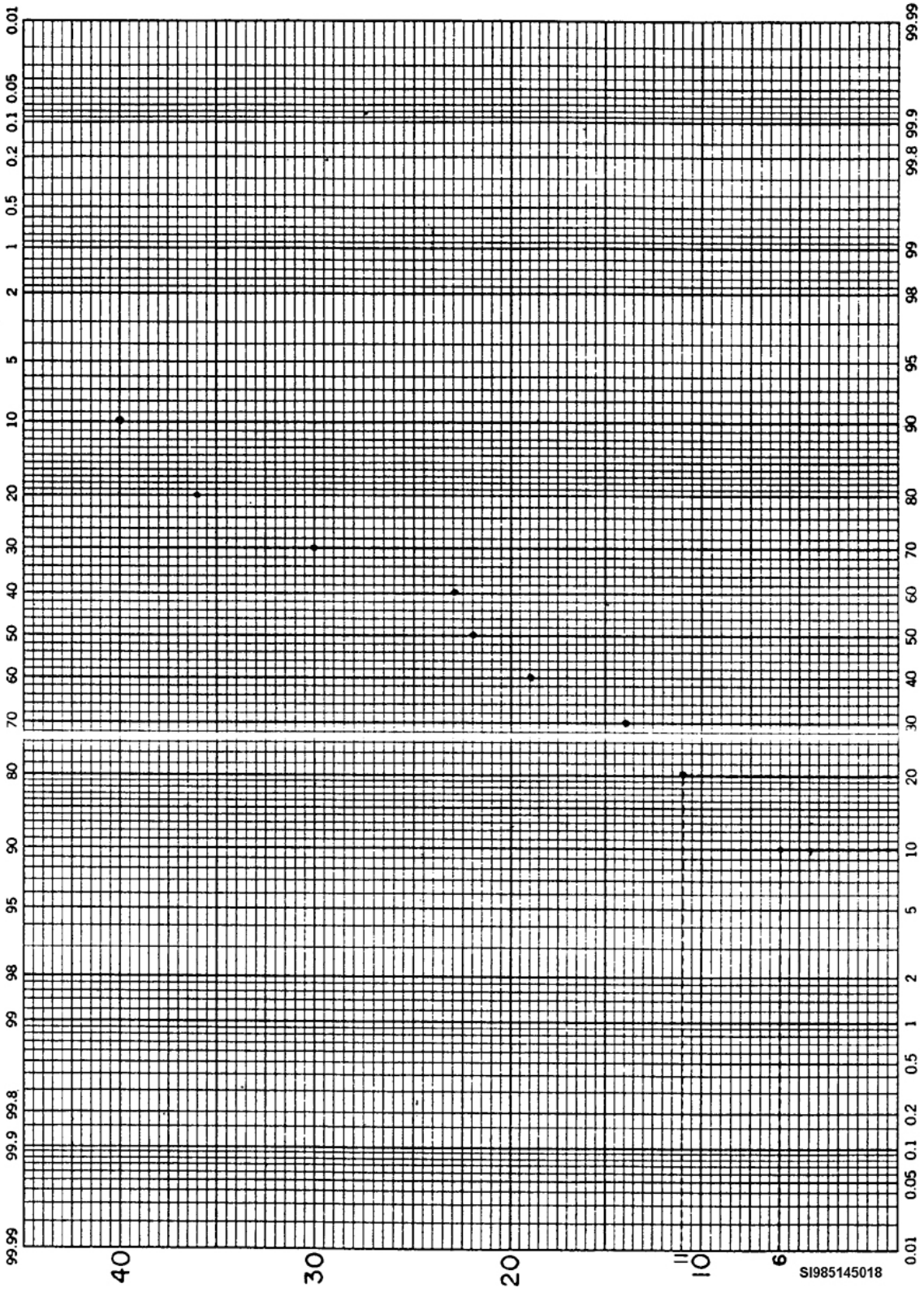


Figure 2–19. Normality plot.

Through practice you will find this method is much easier than it appears to be at first. You will also become more capable of evaluating the degree of normality through continued use. A good method of learning to recognize normal data distributions on graph paper is to plot data known to be skewed and compare its line with normal data, which forms a straight line.

415. Interpreting the normal curve area

Now that you have studied measures of central tendency and measures of variability, you can use these concepts to study the normal distribution in more detail.

Use of the standard deviation

Of all the measures of variability, standard deviation is considered to be the most important. It has many desirable properties and is often used in making further statistical calculations. One of the most interesting properties of standard deviation is its relationship to the normal curve.

When we introduced the normal curve earlier, we briefly described it as symmetrical and bell-shaped. We also told you that the normal curve is a graphical representation of the theoretical normal distribution. Since the normal distribution is symmetrical about its peak in the center, its mean, median, and mode are all located in the center of the distribution and have the same value. An exact percent of the individual values falls within ranges established by the standard deviation in conjunction with the mean. Therefore, when the mean and standard deviation are known, the normal distribution is completely specified. So, 68.26 percent of the individual items are located within a distance of one standard deviation from the mean; 95.44 percent are located within a distance of two standard deviations from the mean; and 99.74 percent are located within a distance of three standard deviations from the mean, as shown by curves A, B, and C, respectively, in figure 2–20.

Once the mean and standard deviation of a normal curve are known, the values of points one, two, or three standard deviations above and below the mean are easily determined. To illustrate, suppose the normal distribution represented by the curves in figure 2–20 has a mean of 10 and a standard deviation of 2. The mean of 10 is recorded at the center of the curve. The point's one standard deviation above the mean and one standard deviation below the mean, plotted on curve A, are determined as follows:

$$\bar{X} + s = 10 + 2 = 12$$

$$\bar{X} - s = 10 - 2 = 8$$

Figure 2–20. Relationships of standard deviation to the normal curve.

Points two and three standard deviations above and below the mean, plotted on curves B and C, are determined in a *similar* manner:

$$\bar{X} + 2s = 10 + 2(2) = 14$$

$$\bar{X} - 2s = 10 - 2(2) = 6$$

$$\bar{X} + 3s = 10 + 3(2) = 16$$

$$\bar{X} - 3s = 10 - 3(2) = 4$$

While the curves in figure 2–20 seem to extend from a point three standard deviations below the mean to a point three standard deviations above the mean, the theoretical normal distribution actually extends infinitely on each side of the mean. In actuality, however, three standard deviations on either side of the mean includes just about all of the cases.

Normal curve area table

All of the percentages shown in figure 2–20 are based on complex mathematical calculations beyond the scope of this text. However, you can quickly and easily arrive at these percentages by using the normal curve area table at figure 2–21. The full table is shown at foldout 1, A.

An important variable in dealing with a normal curve area is the *Z score*. The Z score measures how *many* standard deviations a value is *from* the mean. The normal curve area table contains a four-decimal place listing of proportions of the area under the normal curve. These are the proportions of the area between the mean and a point Z standard deviations from the mean.

NORMAL CURVE AREA TABLE

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990									
3.2	.4993									
3.3	.4995									
3.4	.4997									
3.5	.4998									
4.0	.4999									

DEVELOPED BY THE DEPARTMENT OF AIRCRAFT MAINTENANCE TRAINING,
CHANUTE AFB, IL

Figure 2–21. Normal curve area table.

In figure 2–21, notice that the values of Z are in the left column and across the top. The left column shows a whole number and the first decimal, whereas the row on the top is for the second (hundredths) decimal place. In the table, the number of standard deviations from the mean varies from 0.00 to 4.00.

The proportions in figure 2–21 vary from .0000 to .4999. In terms of percent, this would be from 00.00 to 49.99 percent. This reflects the fact that as you move farther from the mean, you come closer to including 50 percent of the area under half the normal curve and, therefore, 50 percent of the data. The table gives area values for only half or 50 percent of the normal curve in order to conserve space. However, since the normal curve is symmetrical, it is a simple process to convert from the area between the mean and a point on the right to the area between the mean and a point on the left.

NOTE: Figure 2–21 gives numerical distance from the mean to the tails.

Many times when you are solving normal curve area problems, you'll need to find the percent of the area between the mean and specific values in the distribution. When the Z values or Z scores of the points are known, use the normal curve area table to determine the percentages. To determine these Z scores when they are not given, use the following formula:

$$Z = \frac{X - \overline{X}}{s}$$

Where:

Z = number of standard deviations of X from $\overline{X}$.

X = specific value in the distribution

$\overline{X}$ = mean

s = standard deviation

The formula determines the actual distance of a specific point in the distribution from the mean, X – $\overline{X}$, and then divides this distance by the value of a standard deviation. The result of this division is Z, the number of standard deviation units you must measure off from the mean to reach the specific point (X). For example, if 40 is a specific value in the distribution, 30 is the mean value, and the standard deviation is 10, then

$$Z = \frac{40 - 30}{10} = \frac{10}{10} = 1.00$$

The value 40 is 1.00 standard deviation above the mean of 30. Therefore, Z is 1.00. The sign of Z indicates on which side of the mean the specific value is located. If Z is positive, X is to the right of the mean, whereas Z is negative if X is on the left side of the mean. In the normal curve area table, there are no values for negative Z scores. As far as areas go, equal Z scores (whether positive or negative) include equal areas when taken from the mean.

Using the normal curve area table (fig. 2–21), check the percentages in figure 2–20. Figure 2–20, part A, shows the limits of $\overline{X} \pm 1s$ (read the mean plus and minus one standard deviation). Located within these limits are 68.26 percent of the area under the curve and, therefore, 68.26 percent of the data. In this example, where $\overline{X}$ = 30 and s = 10, $\overline{X} \pm 1s$ results in $\overline{X}$ + 1s = (30 + 10) or 40 and $\overline{X}$ – 1s = (30 – 10) or 20. Since 40 is 1.00 standard deviation from the mean, the Z score is 1.00. Go down the Z column in figure 2–20 until you get to 1.0. To the right of 1.0 are 10 numbers. Since the hundredths position in 1.00 is .00, move to the number falling under the heading of .00. This is the first number in the 1.0 row and is .3413 or 34.13 percent. This is the percent of the data between $\overline{X}$ and $\overline{X}$ + 1s or

between 30 and 40. In the same way, 34.13 percent of the data are between $\bar{X}$ and $\bar{X}$ – 1s or between 30 and 20. If you are using a normal curve with $\bar{X}$ = 30 and s = 10, then 34.13 + 34.13 percent or 68.26 percent of the individuals fall between $\bar{X}$ ± 1s or 20 and 40.

Looking back to figure 2–20, parts B and C are explained in a similar manner. Part B involves two standard deviations on each side of the mean. In this case the Z score is 2.00. Looking up 2.00 in figure 2–21, note the proportion of .4772 or 47.72 percent. This is the percent of the area on one side of the mean only, between the mean and point 2.00 standard deviations away. Another 47.72 percent of the area is located on the other side of the mean within two standard deviations. The total percent within $\bar{X}$ ± 2s, then, is 47.72 + 47.72 percent or 95.44 percent, as indicated in part B (fig. 2–20).

Part C (fig. 2–20) shows three standard deviations on each side of the mean. This time Z is 3.00. Looking up 3.00 in figure 2–21, note that .4987 or 49.87 percent of the area under a normal curve is within three standard deviations of the mean on each side of the curve. The total percent within $\bar{X}$ ± 3s then is 49.87 + 49.87 percent or 99.74 percent, as shown in part C (fig. 2–20).

You can also think of standard deviation as an expression in terms of baseline values of the distance in which a given percent of individuals falls from the mean. Because of this characteristic, you will soon learn how to describe data by determining where various percentages of the individuals are located. From the example given s = 10 and $\bar{X}$ = 30, let's work through some typical problems:

1. What percent of the data lies above the value of 45?

$$Z = \frac{45 - 30}{10} = \frac{15}{10} = 1.5$$

   1.5Z = 43.32 percent (from the normal curve area table, fig. 2–21) + 50 percent below the mean. This equals 93.32 percent, which represents 45 or less, leaving 6.68 percent of the cases greater than 45.

2. How many cases and what percent of the values fall between 15 and 25 if our entire population contains 150 cases:

$$Z = \frac{15 - 30}{10} = \frac{-15}{10} = -1.5 = 43.32 \text{ percent}$$

(from normal curve area table, fig. 2–21)

$$Z = \frac{25 - 30}{10} = \frac{-5}{10} = -0.5 = 19.15 \text{ percent}$$

43.32 - 19.15 = 24.17 percent of 150

which = 36.25 cases

3. To determine values from percentages, such as, "What value on the baseline has 90 percent of the data falling below?" We are looking for the value where 90 percent of the curve area is covered from the beginning of the low end tail of the curve towards the positive side of the curve. Since the curve area already covers 50 percent from the low tail end to the mean, we only need to look for 40 percent more above the mean. We start at 90 percent. Then, 90 percent minus 50 percent = 40 percent on the high side of the curve. Looking at the body of the table (fig. 2–19), find the closest value to 40, which is 39.97 and equals a Z value of 1.28. Multiply the s value of 10 times the 1.28Z to get 12.8. So 30 (our mean at 50 percent) plus 12.8 (1.28 standard deviation at 39.97 percent) equals the value of 42.8, which means that only 10 percent of our data is greater than 42.8.

416. Symmetrical and nonsymmetrical curves

This lesson discusses how to distinguish among the various types of symmetrical and nonsymmetrical curves.

Symmetrical curves

If you recall, the histogram and frequency polygons shown previously gave you a general impression of the shape of the data distribution that is portrayed.

When collecting maintenance data, some data curves occur quite frequently, while others are rarely encountered. When discussing these general types of curves, use smooth curves to describe them. All curves that have right and left halves alike, with respect to the center, are called *symmetrical curves*. The two types of symmetrical curves we discuss are normal and rectangle.

Normal curve

The *most* important of the symmetrical curves is the normal curve (fig. 2–22 [1]). Many of our statistical techniques are based on the assumption that the data is from a normal distribution. The normal curve is bell-shaped, with its mode (most frequent value) occurring at the center of the distribution. It is a graphical representation of the normal distribution, which is a theoretical distribution.

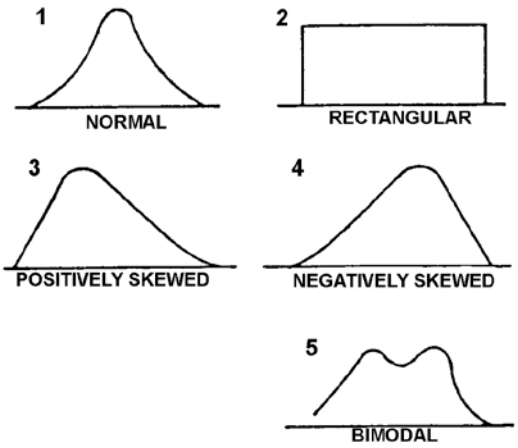


Figure 2–22. Types of data curves.

Rectangular

If a frequency distribution has the same frequency in each class, the histogram has equal height at all points and the distribution is said to be a rectangular distribution. A curve representing a rectangular distribution is shown in figure 2–22 (2). In working with maintenance data, you may occasionally find a distribution that has approximately the same number of frequencies in each class. In such situations, the data fit a rectangular distribution. Also, you may want to make use of a theoretical distribution that is rectangular.

Types of nonsymmetrical curves

A curve that has its right half shaped differently from its left half is called *nonsymmetrical*. Such a curve is said to be *skewed* in the direction to which it extends farthest from the highest point. The three types of nonsymmetrical curves are positive, negative, and bimodal.

Positive skew

After constructing several frequency polygons from actual maintenance data, you may notice that the frequencies often tend to pile up on the left-hand side of the graph with a long tail extending to the right. Such a curve (fig. 2–22 [3]) is said to be positively skewed, or skewed right. This type of curve is very common in portraying maintenance data because a distribution of man-hours required to

complete a unit of work is limited to the left by zero time while on the right there is no such limit. The few jobs that take an extremely long time always appear out in the right tail. Although most curves representing maintenance data tend to be positively skewed, you may occasionally find one skewed in the opposite direction.

Negative skew

When frequencies pile up on the right side with the long tail extending to the left, the curve is negatively skewed or skewed left. Figure 2–22 (4), displays an example of a negatively skewed curve.

Bimodal

Another type of curve that is usually skewed but may be symmetrical is a curve with frequencies piled up at two different places. A frequency distribution that has two classes with high frequencies separated by classes with lower frequencies is called a bimodal distribution. A bimodal curve (fig. 2–22 [5]), has two modes. It is also possible for a curve to have more than two modes. Bimodal distributions often result from mixing data from two different distributions that should be studied separately.

Self-Test Questions

After you complete these questions, you may check your answers at the end of the unit.

412. Computing the standard deviation

1. Define standard deviation in mathematical terms.

2. How are the standard deviation and the variance related?

3. In the equations for calculating the standard deviation, what step gets rid of the negative values?

4. In a population, if all the values are zero, what is the standard deviation?

5. In a normal distribution, as n grows smaller, what happens to s?

413. Computing the standard error of the mean

1. Define the sampling distribution of the mean.

2. The variability of the $\bar{X}$ values in the sampling distribution of means is affected by what two factors?

3. Define the standard error of the mean.

4. The standard error of the mean measures the amount of error between a sample mean and what?

414. Defining normal distribution curves

1. The normal distribution is composed of what two parameters?

2. Like all continuous distributions, the total area under the curve is what percentage?

3. For normality testing, what is the first step in plotting data?

4. What is a good method to learn to recognize normal data distributions?

415. Interpreting normal curve area

1. What percentage of individual items is located within a distance of one standard deviation from the mean in a normal distribution?

2. What does the Z score represent?

3. Given s = 5 and $\bar{x}$ = 20, what percent of data lies *below* the value of 25? *Above* 25?

416. Symmetrical and nonsymmetrical curves

1. What is a symmetrical curve?

2. What is a nonsymmetrical curve?

3. What type of curve is *very common* when portraying the distribution of maintenance data?

2–5. Hypothesis Testing

Hypothesis testing procedures have long been used by research investigators to make decisions about populations on the basis of information from samples. The same procedures can be used in certain analyses situations. In this section, we'll consider the meaning of terms and concepts associated with hypotheses testing, how to formulate hypotheses, and the steps in the hypothesis testing procedure. So, let's discuss developing a hypothesis as it applies to your job.

417. Developing a hypothesis

As an analyst, you will investigate events or problems based on observations or occurrences within maintenance data. There will be questions that need to be answered and you will have an assumption about those questions. You will need to test your assumptions. A *hypothesis* is an assumption to explain an observation or occurrence that can be tested for further investigation. We will show you how to develop the correct hypothesis that will satisfy the conditions of the applicable tests following these three steps:

1. Ask a question.
2. Gather background information.
3. Develop the hypothesis.

The objective of a statistical test is to test a hypothesis concerning the values of one or more population parameters. You will generally have an assumption about the parameter or parameters that you wish to support. For example, you might ask is the mean life (μ_1) of a new type of aircraft tire is greater than the mean life (μ_2) of the old tire. Although, a question comes first in developing a hypothesis, it is presented in the form of a statement. A good hypothesis is specific, has a relationship between the variables and is testable.

Scientific process

Two very important parts of the duties of an analyst are to review maintenance data and identify areas that need further study or investigation. Since there are a large number of items available for study, you will not have time to make a detailed investigation of everything. So, your efforts should be directed to those areas where problems are most likely to exist and detailed investigations are most likely to pay off. The situation would be ideal if you could always decide to look for problems that exist and not look for problems where they do not exist. But, often, this is too much to hope for. When an apparent difference in the performance of like personnel or equipment being employed in the same manner is observed, you are faced with a decision.

The first step is to ask a question. What is causing the difference? Why is there a difference? Does this observed difference represent a real difference or is it merely a chance variation? When is the problem occurring? How is the problem occurring? Should you worry about it? Where is the problem? Does a problem exist? Why is there an issue? Should a more detailed investigation be made?

The second step is to gather background information on the topic. You can begin with the questions asked as a basis. If you do some basic research, it will prevent you from wasting time doing an investigation on an incorrect assumption about the problem. If the apparent difference is very large, you might (by using common sense) decide that a problem does exist and that a more detailed study is needed. On the other hand, if the observed difference is very small, you might decide to forget it and go on to larger problems. But, what if the observed difference is neither very large nor very small? You need a statistical technique to help you decide which way to go. This is one situation where a hypothesis and a hypothesis testing procedure will help you.

The third step is to develop a hypothesis (make an assumption) about what you think is happening based on what you have already learn from your research. Then, by using the proper statistical technique, test the hypothesis and make a decision. If the test fails to meet a certain condition, reject

the hypothesis and state, with a certain degree of confidence, that there is no significant difference. However, if the test meets a certain condition, accept the hypothesis and assume that the difference is, in fact, significant. The conditions we are talking about are prescribed numerical values. These numerical values are called critical values.

This explanation roughly describes the procedure of hypotheses testing. We will explain the meaning of the terms and concepts involved, and give a more detailed description of the procedure. Hypothesis testing then is used in analysis to identify significant differences, trends, and relationships. This testing provides information and helps you, the analyst, decide whether a problem exists and whether a more detailed investigation should be made.

Types of hypotheses

Hypotheses can be stated in several forms, but basically there are two types of hypotheses that you need to know—*null hypothesis* and its *alternate hypothesis*. Another name for the null hypothesis is the *given hypothesis*. In the following hypothesis statements, note the phrase *significant difference*; it implies that the difference between the random samples is due to a definite (assignable) cause and not just random factors.

Null hypothesis

The null hypothesis is phrased in a way that tells you what to expect by giving a specific value to work with. The null hypothesis is symbolized as H_0. It may reflect your basic or preconceived assumptions about the situation or data in question. The null hypothesis assumes that everything is fine, there is no problem, and there is *no* significant difference between two or more population samples. This is the hypothesis that is actually tested.

Example of a null hypothesis:

> H_0: There is no significant difference in the population means of man-hours required to do the job for workers A and B, or in symbolic terms: $\mu_A = \mu_B$

Another way of stating this null hypothesis is to say the population mean of worker A is equal to the population mean of worker B. The null hypothesis gives you a specific value of a population parameter on which to base your expectations. The appearance of an equals sign (=) in a symbolic hypothesis statement reflects this expectation.

Alternate hypothesis

This is the hypothesis that will be of most interest to you. It is the hypothesis you will accept if the null hypothesis is rejected. Normally, the alternate hypothesis is what you believe to be true (i.e., that there *is* a significant difference between two or more population samples). An alternate hypothesis may be denoted by the symbol H_1 or any symbol that best describes its content.

An example of an alternate hypothesis for the null hypothesis, which was given above, may be stated as follows:

> H_1: The population means of man-hours required to do the job for workers A and B are significantly different, or in symbolic terms: $\mu_A \neq \mu_B$.

An alternate hypothesis statement may be directional or nondirectional. The example above is a nondirectional alternate because it doesn't state the direction of the difference (higher or lower). It merely says that the two means are different. A directional alternate states the direction of the difference.

Example of a directional alternate hypothesis:

> H_1: The population mean of man-hours of worker A is significantly greater than the population mean of man-hours of worker B, or in symbolic terms: $\mu_A > \mu_B$.

You will see later that the type of alternate hypothesis you use depends on the nature of the situation and the type of test used.

Sampling distribution

The sampling distribution is a picture of what could happen by chance if the values were truly random. It is a theoretical distribution including all the values that a statistic can take when computed from random samples of equal size. We previously discussed the sampling distribution of the mean in which the sampling distribution was normal. In other cases, it may be a different kind of distribution. By knowing what the sampling distribution of the statistic is like, you can make probability statements about the occurrence of certain values of the statistic. This distribution will be in the form of a table that is used as a comparative base for the test statistic.

418. Level of significance

As an analyst, you must determine the probabilities of arriving at incorrect determinations when formulating your hypotheses and setting up your testing procedures. You do this by setting the level of significance, thereby minimizing the chance of making an error.

Statistical errors

Whenever a hypothesis is accepted or rejected on the basis of information from a sample, there is always the chance of making a wrong decision; that is, rejecting H_0 when H_0 is actually true or accepting H_0 when H_0 is really false.

Possible outcomes

When testing the null hypothesis, you are trying to determine if it is true or false. Since you are usually dealing with *sample* data, and not the whole population, you cannot be absolutely sure that you'll make the right decision. There are four possible outcomes:

1. H_0 is true and you correctly determine that it is true.
2. H_0 is true and you incorrectly assume it is false.
3. H_0 is false and you correctly determine that it is false.
4. H_0 is false and you incorrectly assume it is true.

In the first and third outcomes, a correct decision was made. In the second outcome, a true null hypothesis was rejected. This is a *type I error*. In the fourth outcome, one failed to reject a false null hypothesis. This is a *type II error*. The probability of making a type I error is denoted by *alpha* (α). The probability of making a type II error is denoted by *beta* (β).

Reducing errors

Any time you make decisions in the face of uncertainty, you cannot eliminate the possibility of making a mistake. For tests of hypotheses to aid in good judgment, they must minimize errors of decisions. This is not a simple matter because an attempt to decrease one type of error may increase the chances of another type of error. In practice, one type of error may be more serious than the other; therefore, a compromise should be reached to eliminate the more serious error. The only way to reduce both types of errors is to increase the sample size, which may not be possible or practical. However, you can reduce your chances of making a type I error.

Significance levels

How large a risk of an error are you willing to accept? Five percent? One percent? Whatever the risk factor you are willing to accept, this is your *level of significance*. You use the α (type I error probability) to set up this significance level. Recall from the normal curve area lesson that you can divide the area under the curve into proportions. When you assign a level of significance to a statistical test, you are actually assigning a certain proportion of data in the normal curve area as your *rejection region*. This means that if your test result falls anywhere in this rejection region, you make the decision to reject the hypothesis you are testing.

The reasoning employed in a statistical test of a hypothesis runs counter to our everyday way of thinking; that is, it is similar to the mathematical method of proof by contradiction. The hypothesis that the analyst wishes to "prove" (i.e., support) is the alternative hypothesis. To do this, the analyst tests the converse (opposite) of the alternative hypothesis, the null hypothesis, hoping that the data will support its rejection. Rejection of the null hypothesis implies support for the alternative hypothesis, which was the objective.

Why employ this reverse type of thinking—gaining support for a theory by showing that there is little evidence to support its converse? Why not test the alternative hypothesis? The answer lies in the problem of evaluating the probabilities of incorrect decisions.

If the alternative hypothesis is true, the sample data tends to support rejection of the null hypothesis, that is, the test result will fall within the rejection region. Then, the probability of making an incorrect decision is readily at hand. It is a probability that was specified in setting up the rejection region. Thus, if you reject the null hypothesis (which is what you hope will occur), you immediately know the probability of making an incorrect decision. This gives you a measure of confidence in your conclusion.

Taking the opposite track, if the alternative hypothesis is true, the test statistic will most likely fall in the acceptance region (also called the fail to reject region). Now, to find the probability of an incorrect decision, you must evaluate β, the probability of accepting the null hypothesis when it is false. For most statistical tests, however, it is very difficult to calculate β.

419. Hypothesis testing procedure

The terms and concepts we just discussed will mean more to you as we use them in discussing the procedure for testing hypotheses. Hypothesis testing provides a basis for making a decision that is more reliable than a professional guess, refines the common-sense approach, and attaches a probability to the decision. The procedure for testing hypotheses can be divided into six steps.

1. State the null and alternate hypotheses.
2. Choose the statistical test.
3. Set the level of significance.
4. Determine the sampling distribution.0
5. Compute the test statistic.
6. Make the comparison and decision.

Step 1. State the null and alternate hypotheses

The first step in this decision-making procedure is to state the null and alternate hypotheses. For example, suppose the question is raised as to whether groups A and B differ in the time required to do a job. Since you do not suspect that their population means are different, you make the statement that there is no difference in their population means, and you use it for the null hypothesis. For the alternate hypothesis, it states that their population means are different.

Then, the null and alternate hypotheses could be stated as follows:

H_0: There is no significant difference in the population means of time required to do the job for groups A and B. (This is your testable hypothesis.)

H_1: The population means of groups A and B are significantly different. (If the null hypothesis is rejected, the alternate is available for acceptance.)

Step 2. Choose the statistical test

Once you have stated the hypothesis, your next step is to choose a statistical test. Since there are many tests available, there has to be some rational basis for choosing a specific test. To choose an appropriate test, consider such things as the purpose, power, basic assumptions, and measurement requirements of the test and also the nature of the data to be analyzed. You should look at the following items as suggestions in choosing a test:

Power of the test

Statisticians refer to the value of $1 - \beta$ as the power of a test. This is the measure of how good the test is at rejecting a false null hypothesis. The more "powerful" a test is (the closer the value of $1 - \beta$ is to 1), the more likely the test is to reject a false null hypothesis.

Classification of data

All tests are not applicable to all types of data.

Type of test

There are two types of tests available: parametric and nonparametric. Each has certain requirements or prerequisites to comply with before using the test. These requirements or prerequisites are called the *basic assumptions* of the test.

Sample size

The sample size can be a limiting factor with certain tests. The test you choose may require a sample size greater than 30, require a sample size 30 or less, or set no sample size prerequisites. This choice is often determined by the amount of time, availability of data, and ease of computation.

Step 3. Set the level of significance

Having stated the hypotheses and selected the statistical tests, the next step is to set the significance level (α). You are basically determining how large a risk you are willing to take in making a type I error. Remember, however, that the smaller the significance level α you choose, the greater the possibility of making a type II error. This step is optional at this point and may be performed later as part of the decision step. By presetting the level of significance, you make the decision more objective with less chance of doubt attached to it.

When the level of significance is set in advance, it is selected arbitrarily, and the values of .05 or .01 are often used. These are the generally accepted standards. We will use the values of .10, .05, or .01 in the discussions that follow. In practice, however, there are times when you may want to use levels of significance that are smaller to reduce the chance of making a type I error. As you gain experience, there will be times when you simply know at what level of significance the hypothesis could be accepted or rejected.

Step 4. Determine the sampling distribution

After setting the level of significance, your next step is to determine the sampling distribution to be used as the basis for comparison. Make your comparison indirectly by going to tables of probability or tables of critical values, which have been developed to make your job easier. For example, if the sampling distribution is normal, use the normal curve probability table. If the sampling distribution is made up of U values, as in the Mann-Whitney U test, use the table of critical U values, and so on.

Step 5. Compute the test statistic

The next step in the hypothesis testing procedure is to compute a test statistic. You compute the test statistic using formulas associated with a particular statistical test. This step is usually the most time-consuming and may require several calculations. Do not become discouraged if the calculations are tedious. It i's far better to spend a couple hours to determine that further research is unnecessary than to spend two weeks of research only to find that a problem does not exist.

Step 6. Make the comparison and decision

The last step is making a comparison and a decision. After you have computed the test statistic, compare it to the critical value of the sampling distribution as determined by the level of significance. This procedure lets you measure the probability that your null hypothesis is true. Confidence in your decision is based on the level of significance used in the comparison.

Self-Test Questions

After you complete these questions, you may check your answers at the end of the unit.

417. Developing a hypothesis

1. List the three steps in the development of a hypothesis.

2. Why detailed investigations cannot be made on everything in maintenance?

3. Why is it important to do some basic research before creating a hypothesis?

4. What is the purpose of hypotheses testing in analysis?

5. How is the null hypothesis denoted?

6. Define null hypothesis.

7. Define alternate hypothesis.

8. Why are sampling distributions used in hypotheses testing?

418. Level of significance

1. When a given hypothesis is rejected, when in fact it is true, what type of error has been made?

2. What does alpha (α) denote?

3. Define level of significance.

4. What is used to set up the level of significance?

419. Hypothesis testing procedure

1. List the six steps of hypothesis testing procedure.

2. State an advantage of presetting the level of significance before making the comparison step in hypothesis testing.

3. Which step in the hypothesis testing procedure normally takes longer to perform? Why?

4. Which factor of hypothesis testing determines the level of confidence in your decision?

Answers to Self-Test Questions

405

1. Discrete data.
2. The total set of data.
3. Measures that characterize a population.
4. A part of a population or a part of the whole.
5. A sample where each item in the population has an equal chance of being included in the sample chosen.
6. Dividing the population into subgroups (strata) in such a way that there is as great a homogeneity as possible within each stratum and as great a heterogeneity as possible between each stratum.
7. (1) Unintentional bias occurs because of insufficient planning.
   (2) Purposeful bias is used when you desire data for comparison studies that must meet certain qualifications, such as specific temperature ranges, skill levels, time periods, and so forth.

406

1. (1) Interval.
   (2) Ratio.
2. Nominal.

3. (1) c.
   (2) a.
   (3) b.
   (4) d.

407

1. To summarize data and describe variation in performance.
2. Determine the range of the data.
3. 10 to 20.
4. You lose smoothness and simplicity and are left with a ragged distribution.
5. Displays the number of values falling above or below a certain point on the measurement scale.
6. By dividing the individual class frequency by the total frequencies involved.
7. 12 percent.
8. 42 percent.
9. 68 percent.
10. 4 replacements.

408

1. The assumption that individual values falling within each interval are evenly distributed over the interval.
2. With the corresponding midpoints.
3. Frequency polygons.

409

1. The most frequently occurring value in a distribution.
2. (1) It is the most typical value.
   (2) It is not affected by extreme values.
   (3) It is simple to estimate.
   (4) It is very unstable.
   (5) It is the most appropriate measure that can be used with data from a nominal scale.
3. By finding the value that occurs most frequently.
4. 8.
5. By finding the class with the highest frequency.
6. By drawing a line perpendicular to the baseline from the point where the dotted lines cross.

410

1. The center value in a distribution with half of the individual values on either side.
2. (1) It divides the distribution into two equal parts.
   (2) Any randomly selected value may fall either above or below the median.
   (3) Each value must be arranged according to size before the median can be found.
   (4) It is affected by the total number of items.
   (5) It is not as familiar as the mean.
   (6) It can be used only with data from ordinal, interval, or ratio measurement scales.
3. The middle value is taken as the median.
4. 6.
5. The median is assumed to lie halfway between the two middle values.
6. 6.5.

411

1. When you wish to place equal emphasis on each value in the data distribution.
2. That point in a data distribution about which the sum of the deviations equals zero.
3. (1) The sum of the deviations from the mean is zero.
   (2) The sum of the squares of the deviations from the mean is less than about any other point.
   (3) The value of the mean is determined by every item in the distribution.
   (4) It's greatly affected by extreme values.
   (5) It can be used with data from interval or ratio scales. It should not be used with data from nominal or ordinal scales.
4. 7.
5. Averaging rates, particularly rates of time.
6. By multiplying an item's value in the series by its appropriate quantity factor.
7. If zeros are present in the series.

412

1. The square root of the squares of all the deviations from the mean. It is the actual deviation or distance between every individual value from the mean.
2. The standard deviation is the square root of the variance.
3. Squaring the deviations.
4. Zero.
5. It becomes less representative of the population.

413

1. A theoretical frequency distribution of mean values.
2. The variability of the population from which the random samples were drawn and the random sample size.
3. The standard deviation of a sampling distribution of means that is equal to the standard deviation of the population divided by the square root of the sample size.
4. The true population mean.

414

1. The mean and the standard deviation.
2. 100 percent.
3. Arrange the data in an array from the lowest to the highest value.
4. Using graph paper, plot data known to be skewed and compare its line with normal data, which forms a straight line.

415

1. 68.26 percent.
2. How many standard deviations a value is from the mean.
3. 84.13 percent; 15.87 percent.

416

1. All curves that have right and left halves alike, with respect to the center.
2. A curve that has its right half shaped differently from its left half.
3. A positively skewed or skewed right curve.

417

1. (1) Ask a question.
   (2) Gather background information.
   (3) Develop the hypothesis.
2. Due to the large number of items available for study.
3. To prevent wasted time on an incorrect assumption about a problem.

4. To identify significant differences, trends, and relationships.
5. $H_{0.}$
6. Your basic or preconceived assumptions about the situation in question.
7. The hypothesis available for acceptance if the null hypothesis is rejected.
8. To make probability statements about the occurrence of certain values of the statistic.

418

1. Type I.
2. The probability of making a type I error.
3. The risk factor you are willing to accept.
4. The assignment of a certain portion of data in the normal curve area as your rejection region.

419

1. (1) State the null and alternate hypothesis, (2) choose the statistical test, (3) set the level of significance, (4) determine the sampling distribution, (5) compute the statistic, and (6) make the comparison and the decision.
2. It makes the decision more objective with less chance of doubt attached to it.
3. Step 5; computing the statistic requires several calculations.
4. The significance level used in the comparison.

Complete the unit review exercises before going to the next unit.

Unit Review Exercises

Note to Student: Consider all choices carefully, select the *best* answer to each question, and *circle* the corresponding letter. When you have completed all unit review exercises, transfer your answers to the Field-Scoring Answer Sheet.

Do not return your answer sheet to AFCDA.

11. (405) Which is the *best* example of a population in statistical language?
    a. Yearly B–1 failures at Dyess AFB.
    b. Weekly C–130 failures at Scott AFB.
    c. Monthly E–3C failures at Tinker AFB.
    d. Total T–38s failures across the Air Force.

12. (405) A sample of a population that is taken in such a manner that *each* value has an *equal* chance of being selected is referred to as a
    a. biased sample.
    b. random sample.
    c. sampling theory.
    d. sampling application.

13. (405) If you construct a query language processor (QLP) retrieval to select every 8th record, you are using which sampling technique?
    a. Selective.
    b. Stratified.
    c. Systematic.
    d. Purposeful.

14. (406) Which measurement scale consists of *equal* intervals between scale values and an *arbitrary* zero point?
    a. Ratio.
    b. Nominal.
    c. Ordinal.
    d. Interval.

15. (406) You are given two pieces of test equipment that must be loaded on a pallet. One piece weighs 125 pounds and the other piece weighs 3.5 times as much. Using the ratio measurement scale, how much does the second piece of equipment weigh?
    a. 375.0 pounds.
    b. 415.5 pounds.
    c. 437.5 pounds.
    d. 500.0 pounds.

16. (406) Given measurements of 5.0 hours, 10.0 hours, 15.0 hours, and 20.0 hours, what type of data and measurement scale would you use to classify these data items?
    a. Discrete; ratio.
    b. Discrete; interval.
    c. Continuous; ratio.
    d. Continuous; interval.

17. (407) With a noncumulative frequency distribution range of 3.6, which class interval will give you 18 classes?
    a. 0.1.
    b. 0.2.
    c. 0.3.
    d. 0.4.

18. (407) One way of *comparing* class frequencies to the total frequency is by
    a. indirect comparison.
    b. direct comparison.
    c. visual inspection.
    d. percentage.

19. (408) When constructing a frequency polygon, what are plotted against the corresponding midpoints?
    a. Series of rectangles.
    b. Individual values of data.
    c. Lower limit and baseline.
    d. Frequencies of the various class intervals.

20. (409) The mode is the *only* measure of central tendency that can be used with which measurement scale?
    a. Ratio.
    b. Interval.
    c. Ordinal.
    d. Nominal.

21. (410) Analysts frequently use the median because it is easy to compute and gives a better picture of data than the mean and mode when data are
    a. normal.
    b. skewed.
    c. incomplete.
    d. inconclusive.

22. (410) The median *cannot* be used with data from which measurement scale?
    a. Nominal.
    b. Interval.
    c. Ordinal.
    d. Ratio.

23. (410) What is the median of 11, 2, 5, 12, 14, 16, and 18?
    a. 12.
    b. 13.
    c. 14.
    d. 15.

24. (411) A *true* characteristic of the arithmetic mean is it is
    a. affected by extreme values.
    b. not usable with the ratio measurement scale.
    c. not affected by the number of items in the distribution.
    d. the most frequently occurring value in the distribution.

25. (411) The harmonic mean is used *primarily* for averaging
    a. data of varying weights.
    b. skewed distributions.
    c. approximate values.
    d. rates.

26. (411) What is the arithmetic mean of the values 8, 10, 11, 11, and 5?
    a. 5.
    b. 6.
    c. 9.
    d. 11.

27. (411) Compute a weighted mean for a distribution containing two values of 3 each, four values of 2 each, and four values of 5 each.
    a. 1.0.
    b. 3.4.
    c. 6.6.
    d. 11.3.

28. (411) Three workers perform a similar task. Worker A takes 30 minutes to complete the task and can finish 2 jobs per hour. Worker B takes 20 minutes to complete the task and can complete 3 jobs per hour. Worker C takes 40 minutes to complete the task, and completes 1.5 jobs per hour. Which calculation method will you use to find the average time it takes to complete the job?
    a. Harmonic mean.
    b. Arithmetic mean.
    c. Weighted harmonic mean.
    d. Weighted arithmetic mean.

29. (411) A *unique* feature of the harmonic mean is it
    a. can only be used for grouped data.
    b. will always be less than the arithmetic mean.
    c. is only used to average skewed distributions.
    d. is never weighted when various quantities of the denominator factors are used.

30. (412) For any distribution, the sum of the deviations is
    a. one.
    b. zero.
    c. less than one.
    d. more than one.

31. (412) What is the population standard deviation for the following values: 6, 8, 9, 14, and 22?
    Formula:
    $$\sigma = \sqrt{\frac{\sum (X - \mu)^2}{N}}$$
    a. 5.7.
    b. 6.4.
    c. 11.8.
    d. 32.5.

32. (412) For a standard deviation of a population, if the number of values *increases*, the standard deviation
    a. increases.
    b. decreases.
    c. formula changes.
    d. remains the same.

33. (412) In a sample, you have 10 X-values, and each value is equal to 7. What is the standard deviation of the sample? Formula for standard deviation:

$$s = \sqrt{\frac{\sum \left(X - \overline{X}\right)^2}{N - 1}}$$

    a. 0.
    b. 2.6.
    c. 14.
    d. 49.

34. (413) Given a large number of random samples, the mean of all the sample means related to the population mean is
    a. the same.
    b. radically different.
    c. exactly 3 standard deviations apart.
    d. more than 3 standard deviations apart.

35. (413) Given a large number of random samples, how do mean values vary with *sample* size?
    a. Directly.
    b. Linearly.
    c. Inversely.
    d. Nonlinearly.

36. (413) In estimating the standard error of the mean, for what sample size do you use *n*–1 in the formula?
    a. 10 or more.
    b. 50 or more.
    c. Less than 25.
    d. Less than 30.

37. (413) If $s_{\overline{X}} = \frac{s}{\sqrt{n - 1}}$, what is the estimate of the standard error of the mean of a sample with a standard deviation of 3 and a sample size of 10?
    a. .3.
    b. .33.
    c. 1.0.
    d. 5.0.

38. (414) A *normal* distribution contains what two parameters?
    a. Mean and mode.
    b. Standard error and mode.
    c. Mean and standard deviation.
    d. Standard error and standard deviation.

39. (414) In a normal distribution, how many standard deviations on each side of the mean contain over 99 percent of the area under the normal area curve within?
    a. 1.
    b. 2.
    c. 3.
    d. 4.

40. (414) The *total* area under a *normal* curve is
    a. 50 percent.
    b. less than 50 percent.
    c. between 50 and 100 percent.
    d. 100 percent.

41. (414) When plotted on *normal* probability graph paper, data from a *normal* distribution shows up as a
    a. circle.
    b. curved line.
    c. straight line.
    d. bell-shaped curve.

42. (415) If $\overline{X}$ equals 24 and s equals 6, what are the values of $\overline{X} \pm 2s$?
    a. 24 and 48.
    b. 18 and 36.
    c. 12 and 36.
    d. 12 and 24.

43. (415) How many standard deviations are represented by a value of 22 if $\overline{X} = 14$ and s = 5?

    Formula: $Z = \dfrac{X - \overline{X}}{s}$

    a. –2.8.
    b. –1.6.
    c. 1.6.
    d. 2.8.

44. (416) Where does the *most* frequent value of a normal curve occur?
    a. At the center of the distribution.
    b. At all points in the distribution.
    c. At the tail end of the distribution.
    d. Away from the center of the distribution.

45. (416) Which curve is *nonsymmetrical*?
    a. Skewed.
    b. Normal.
    c. U-shaped.
    d. Rectangular.

46. (417) What are the steps for developing a hypothesis?
    a. Gather background information and write the hypothesis.
    b. Identify the problem, review the data and write the hypothesis.
    c. Ask a question, investigate the problem and write the hypothesis.
    d. Ask a question, gather background information and write the hypothesis.

47. (417) What type of hypothesis assumes *no* significant difference between two or more population samples?
    a. Alternate.
    b. Directional.
    c. Null or given.
    d. Directional alternate.

48. (417) What type of hypothesis assumes there *is* a significant difference between two or more population samples?
    a. Alternate.
    b. Directional.
    c. Null or given.
    d. Directional given.

49. (418) The probability of making a type I statistical error is denoted by
    a. alpha (α).
    b. beta (β).
    c. mu (μ).
    d. rho (ρ).

50. (419) The measure of how good a statistical test is at rejecting a false null hypothesis is called the
    a. distribution of a test.
    b. sample size of a test.
    c. power of a test.
    d. test statistic.

51. (419) When choosing a statistical test, besides the power of the test and the classification of data, you should also consider the sample size and
    a. distribution.
    b. type of test.
    c. z test statistic.
    d. null hypothesis.

Unit 3. Statistical Process Control

THIS UNIT CONTAINS the basic information necessary to understand the theory behind control charts. You'll study five basic kinds of control charts, three control charts for variables and two for attributes.

3–1. Control Chart Theory

Control charts are important tools used to study and control repetitive maintenance performance. They indicate the manner in which a process is operating and when to make corrections in order to maintain quality.

420. Control charts

To understand the theory of control charts, you first must become familiar with the terms used in conjunction with these charts.

Variables and attributes

As you review the various types of control charts, be certain to distinguish between the method of variables and the method of attributes. Under the *method of variables*, quality characteristics, such as man-hours required for inspections, time required to repair certain equipment, mean time between failures, and so forth, are actually measured and quality is said to be expressed as a variable. The method of variables involves a quantitative classification of data.

When the *method of attributes* is being used, items are placed in categories according to some observed quality characteristic. Data collected by the method of attributes consists of the number of items possessing or not possessing certain errors or discrepancies. For example, there may be only two categories, such as accept or reject, go or no-go, pass or fail, and so forth. When it is either impossible or undesirable to measure a given characteristic of an item, use the method of attributes to make a nonquantitative classification.

Chance and assignable causes of variation

In the repeated performance of any maintenance action, a certain amount of variation is expected and unavoidable. Variations in the quality of maintenance are bound to occur. No two pieces of equipment are exactly alike. No two workers are exactly alike, nor are the conditions under which maintenance actions are performed. The causes of variation may be separated into two types—chance causes and assignable causes.

Chance causes

Chance causes of variation are a whole multitude of small influences, which are a natural part of the maintenance process and are not worth looking for. Chance causes of variation are always present and are not readily within our power to identify or eliminate.

Assignable causes

Assignable causes of variation can be identified, regulated, and possibly eliminated. Usually, these causes are intermittent and may arise from specific disorders in the maintenance process or from errors in reporting. Assignable causes are few in number, but each has a marked effect on the pattern of variation. An important part of your job is to identify assignable causes and determine what made them occur.

Control chart general characteristics

A control chart consists of the features shown in figure 3–1. It has a centerline (CL), which represents the average, an upper control limit (UCL), and a lower control limit (LCL). Points representing the quality of performance are plotted in the order of production. The points tend to vary above and below the centerline, as shown in figure 3–1. Where only chance causes of variation are operating, most of the points fall between the upper and lower control limits.

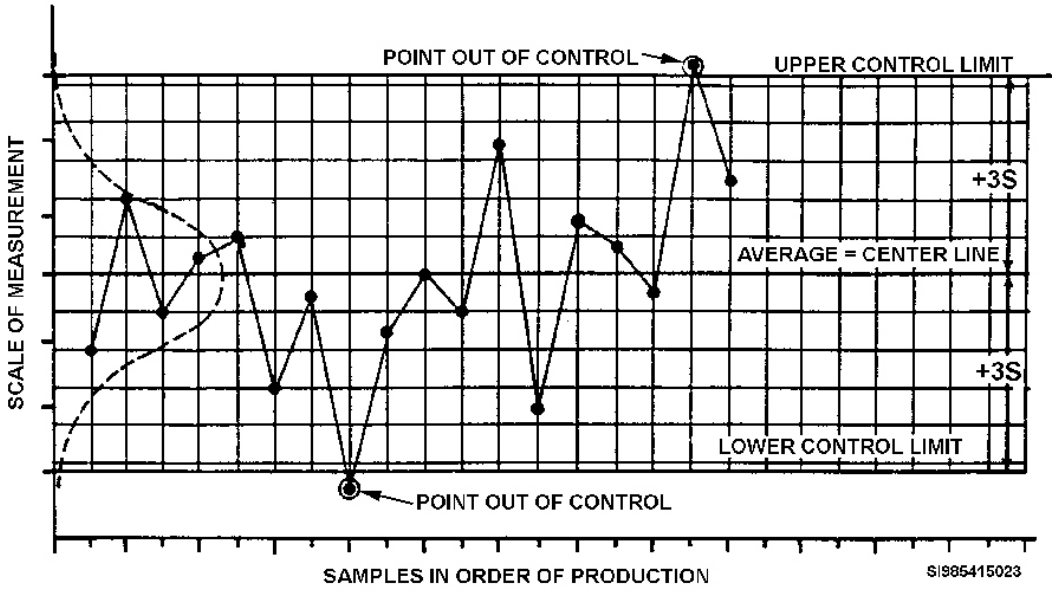


Figure 3–1. Typical control chart.

When a point representing a sample or an individual value falls *outside* the control limits, as shown in figure 3–1, it could have happened by chance. But, the best bet is to say that the extreme variation is due to assignable cause. Points outside the control limits are called out-of-control points.

Maintenance analysts use control charts to detect the presence of assignable causes for variation in a maintenance process. The control chart tells you *when* to look for a problem, but it does *not* identify the problem. When a control chart indicates that a problem exists, you must identify its cause and recommend a corrective action to prevent its recurrence.

421. Determining control limits and out-of-control points

When a control chart is used, it is always hoped that the process under investigation is in control. This is not always the case. A process can be out of control in many ways. If a process is out of control, this does not mean a process is doing badly. Out of control guidelines exist to indicate that a process is not operating within normal parameters. For instance, if a process or rate does drastically better

than normal for a particular month, that point will typically fall outside of the control limits. Using guidelines and setting controls limits will help you determine when a process is out of control.

Out of control guidelines

Guidelines assist in finding when a process is out of control. If the process you are looking at has any of the following indicators, the process is out of control. The out of control guidelines are as follows:

1. Seven successive points on one side of center line.
2. 10 out of 11 successive points on one side of center line.
3. 12 out of 14 successive points on one side of center line.
4. 14 out of 17 successive points on one side of center line.
5. 16 out of 20 successive points on one side of center line.
6. Significant change in average.
7. Significant change in variance.
8. Significant change in variance or average.

Out of control processes must be investigated to determine if the point was due to chance or assignable cause. Determining whether a process is out of control a key role as an analyst and will help you find any major problems within your unit.

Setting control limits

Since control limits are arbitrarily selected, there is always a possibility that the limits may be set too close or too far apart for a particular maintenance situation. When the control limits are set *too close* together, too much time is spent looking for problems that do not exist. This costs time and money and may tend to cause one to lose confidence in control charts. In addition, personnel and equipment may be unjustifiably blamed for deviations, which were actually a normal part of like processes under study. On the other hand, when control limits are set too far apart, the error of looking for trouble that does not exist is rarely made. But this is at the expense of too often failing to look for problems that do exist. In some circumstances, this can lead to serious consequences.

There is a set procedure for establishing control limits. Your first step is to determine the criticality of the item you are evaluating; that is, its importance and the consequence of its failure. Therefore, consider items such as the mission and the safety of both the equipment and the people involved. It may seem ironic, but you may have to compute the control limits twice if your limits are to be the standard for future comparison. During the first computation, include all data. If any out-of-control points exit, discard them and recompute the control limits. The intent is not to include any unusually high or low values when establishing control limit. In other words, the control limits are established for a process that is in control.

However, if you are doing a one-time study to identify a problem area, out-of-control points must be reflected. Otherwise, an elimination of extreme values would defeat your purpose.

Common sense with control charts

Although we used three standard deviation control limits throughout this lesson, in some situations closer limits may be warranted. Remember, with three standard deviations, you expect 99.74 percent of a normal distribution to fall within the control limits. Some processes you analyze will operate so smoothly (without much variation) that you will need to close the control limits to control variation. If you create a control chart with control limits of *three standard deviations* and later find that not enough time is spent looking for assignable causes, *switch to tighter controls*. In situations where the consequences of failing to discover problems are serious, it makes sense to use close control limits. Just remember that when close control limits are used, a larger percent of the points that fall outside the control limits will be due to chance.

Self-Test Questions

After you complete these questions, you may check your answers at the end of the unit.

420. Control charts

1. Define the method of variables.

2. Define the method of attributes.

3. Define chance causes of variation.

4. Define assignable causes of variation.

5. Why do analysts use control charts?

421. Determining control limits and out-of-control points

1. What helps you determine when a process is out of control?

2. Why is an investigation necessary when points fall outside the established control limits on a control chart?

3. What risks are taken when you use control limits that are set too close together on a control chart?

4. What should the analyst do when it has been decided that not enough time is being spent investigating assignable causes for variation?

3–2. Control Charts for Variables

Control charts for variables are excellent for measuring quality characteristics to determine how closely data conforms to standards. There are three control charts you can use when the characteristic to be controlled is measurable—the chart of individuals, the chart of averages, and the chart of averages and ranges.

422. Using charts for individuals

The control chart for plotting individual X values uses the arithmetic *mean* ($\bar{X}$) as its CL and the standard deviation (s) as its measure of variability. To make a control chart for individuals, first gather the raw data representing the maintenance process under study. Next, compute the mean and standard deviation. Then, add the value of the desired number of standard deviations to the mean to determine the UCL, and subtract the same value from the mean to determine the LCL. You must determine the actual number of standard deviations to use, depending on how tight you wish to control the process. The z-score is expressed in terms of the number of standard deviation from the mean. Use the following formula to establish the UCL and LCL:

$$UCL = \bar{X} + (z)s$$

$$LCL = \bar{X} - (z)s$$

Where: z represents the desired number of standard deviations (e.g., 1, 2, or 3).

For example, a study was conducted to find out how much time is being expended in repairing a selected piece of equipment. Three groups of samples were taken, one from each workcenter that repairs the same type of equipment. The time was measured in man-hours. The 30 values in figure 3–2 represent repair times or man-hour expenditures required to repair the equipment under study, which occurred after the control limits were established. These 30 man-hour expenditures are plotted on the control chart in figure 3–3 in the order in which they occurred. The control chart was developed from related data, in this case, a larger but hypothetical group of samples. For our study, the standard deviation and the mean are arbitrary numbers from this large data. The centerline and three standard deviation control limits for this chart were computed from a large sample of data having a mean of 3.2 and a standard deviation of 0.8. We will use the control chart (fig. 3–3) to see how our three samples measure up. The variables under study are man-hours required to repair certain equipment. We plot all the values from the three samples. Therefore, you will see 30 plotted points on the control chart in figure 3–3. Notice the pattern of variation appears to remain within limits until the 25th repair time. Points 25 and 29 are above the upper control limit and the mean shows a rising trend, thus indicating the presence of assignable cause of variation.

Sample 1	Sample 2	Sample 3
5.0	2.4	4.0
2.4	2.0	4.4
3.4	4.0	3.0
2.6	2.8	2.7
4.6	3.9	6.0
4.7	2.2	2.3
2.0	4.8	5.2
3.6	2.6	3.6
1.8	2.8	6.8
4.5	1.8	3.0
$\bar{X}$ = 34.6 ÷10 $\bar{X}$ = 3.46	29.3 ÷10 2.93	41 ÷10 4.1

Figure 3–2. Man-hours required to repair selected equipment.

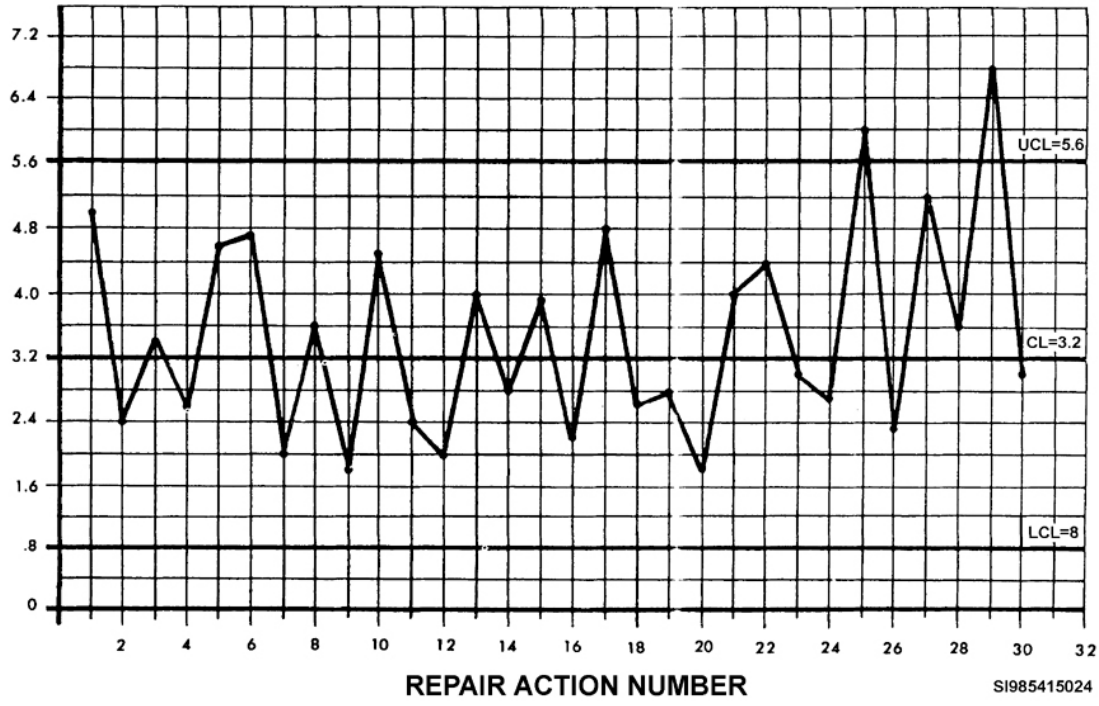


Figure 3–3. Control chart for individuals.

You may use a chart for individuals to study repetitive-type maintenance jobs or inspections, where measurements of time required are made in man-hours or clock hours. Apply it in situations where you do not have enough data to run a chart for averages and in situations where you want to plot every individual value. It works *best* in repetitive-type situations where the distribution of the individual values is fairly symmetrical. If the distribution is *extremely skewed*, the interpretation of a control chart of individuals is *distorted* because the control limits are based on the sampling distribution values from a normal distribution.

423. Using charts for averages

Your *best use* of the control chart for averages is when dealing with a *relatively large* amount of data. You use a control chart for averages when it becomes impractical to plot every value on the chart, as in the control chart of individuals. For example, we took three samples for our chart of individuals, each with 10 values. If we had selected 20 samples of 10 values per sample, then there would have been 200 plotted points on the chart. This is where it becomes *appropriate* to use a chart for averages.

As the name implies, the chart for averages is a control chart for plotting *means of small samples*. When working with means, the sampling distribution is a distribution of means. A chart for averages uses the *same* centerline as a chart for individuals, although it is estimated differently. The measure of variability for the sample means is the standard error of the mean. Calculate the measure of variability using the same centerline as the standard deviation, except the data will be made up of sample means rather than individuals.

To illustrate what a chart would look like, a control chart for averages is shown in figure 3–4. Compare this chart with the chart for individuals shown in figure 3–3. The centerline and control limits of both control charts were computed from the same large sample of data having a mean of 3.2 and a standard deviation of 0.8. Notice that both control charts use the same value (3.2) as a centerline, although the control limits for the chart of individuals were determined by adding to and

subtracting from the mean three standard deviations. The control limits for the chart of averages in figure 3–4 were computed for a distribution of means of small samples of size 10. The standard error was estimated using this formula:

$$s_{\bar{X}} = \frac{s}{\sqrt{n-1}}$$

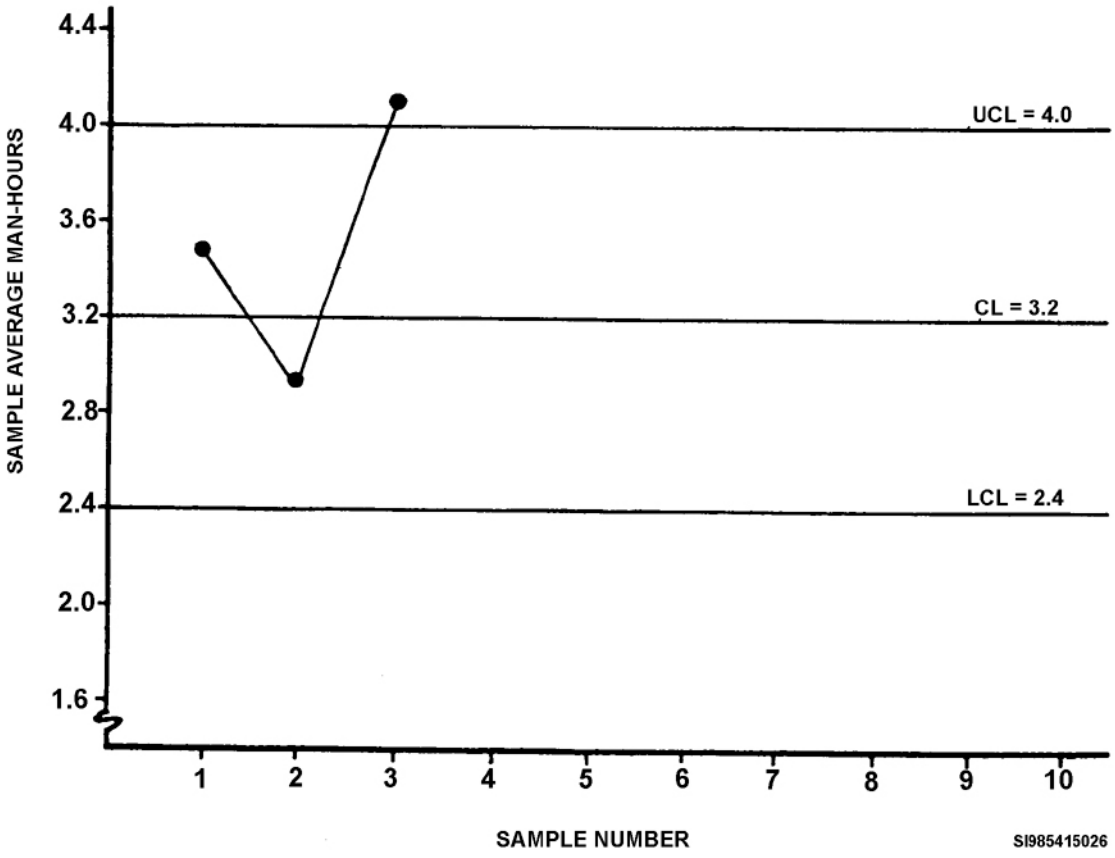


Figure 3–4. Control chart for averages.

Then three standard errors of the mean were added to the centerline value of 3.2 to find the UCL, and three standard errors of the mean were subtracted from the centerline to find the LCL.

For purposes of comparison, let's look at the same three samples from figure 3–2. We will only use the three columns of data in figure 3–2 for our three small samples of size 10. The means or averages (3.46, 2.93, and 4.1) have been plotted on the chart for averages in figure 3–4. In other words, the two control charts (figs. 3–3 and 3–4) are being run on the same maintenance process. Since the sample size in figure 3–4 is 10, there is only one point on the chart of averages for every 10 points on the chart of individuals. If you look at point 3 on the chart of averages, it exceeded the UCL of 4.0. Notice, however, both charts indicate the presence of an assignable cause for variation. When the two X values on the chart for individuals fell above the UCL, the mean of the sample that included those same X values also fell above the UCL on the chart for averages. Look at sample 3 of figure 3–2 and see that one item has a high value (6.8), which drove the average higher than samples 1 and 2. Sample 3 has higher than normal value or values.

Because a distribution of means *tends to be normal*, even when the population is skewed, a control chart for averages has an *advantage* over a chart for individuals. The interpretation of a chart for averages is less likely to be distorted because of the skewness in the population. Like the sampling

distribution associated with a chart of individuals, a chart of averages is based on a normal distribution.

424. Developing charts for dispersion

To describe how values are spread out in a process, use either a range chart (R chart) or a chart for standard deviation. Both charts provide basically the same results and are used with a chart for averages ($\bar{X}$ chart) to control a process. Since the R chart is much easier to compute than the standard deviation, this lesson only covers the R and $\bar{X}$ charts.

Charts for averages *measure changes* in the means of a series. As you have learned, the means may remain constant even with large fluctuations in the "scatter" of values around the mean. If the distribution of the population is normal when sampling methods are used, the distribution of sample means is normal regardless of sample size. In addition, if the population distribution is not normal, the distribution of sample means approaches normality as the sample size increases. Therefore, you may make assumptions about the outcome of means from population samples. Just as each population has its own average and standard deviation, so does each distribution of means standard deviations, or ranges have their own averages and standard deviations.

In this lesson, the ranges are used in establishing the centerline and control limits for the R and $\bar{X}$ charts. Unfortunately, statistical theory cannot give you much useful information about the expected outcome of ranges of samples. However, for a normal distribution, as sample sizes increase, the distribution and ranges become closer and closer to symmetrical.

Because of the infrequent use of R charts in analysis work, calculation of the standard deviation needed for the charts is not taught in this course. Instead, a much simpler formula and a preconstructed table show upper and lower control limits. Refer to figure 3–5, which has been developed for three standard deviation controls, as you study the construction procedures for R and $\bar{X}$ charts. Since R and $\bar{X}$ charts normally are used in conjunction, the table provides data to simplify the computation of the $\bar{X}$ chart.

n	A_2	D_3	D_4
2	1.88	0	3.27
3	1.02	0	2.57
4	0.73	0	2.28
5	0.58	0	2.11
6	0.48	0	2.00
7	0.42	0.08	1.92
8	0.37	0.14	1.86
9	0.34	0.18	1.82
10	0.31	0.22	1.78
11	0.29	0.26	1.74
12	0.27	0.28	1.72
13	0.25	0.31	1.69
14	0.23	0.33	1.67
15	0.22	0.35	1.65

Figure 3–5. Conversion factors for three standard deviation control limits of $\bar{X}$ and R charts.

Use the following steps to construct the R and $\bar{X}$ control charts:

1. Select the characteristic (X) that is to be controlled.
2. Choose a suitable measuring device.
3. Decide on a subgroup (sample) size (n). Subgroups of 4 or 5 are commonly used. However, the size may be any number from two up, depending on the particular situation. Keep the subgroups small (4 or 5) since the essential idea of $\bar{X}$ and R control charts for variables is to select subgroups to keep variation within the subgroup at a minimum. Once you decide on the sample size, all subsequent samples must be of the same size since the sample size sets the control limits.
4. Select a sample size (n) from the process, measure each item in the sample, and record the measurements.
5. Find the mean of the subgroup ($\bar{X}_1$) and its range (R_1) to get your first sample value.
6. Repeat the operations in steps 4 and 5 until you have measured about 25 subgroups. You now have 25 averages (25 $\bar{X}$s), and 25 ranges (25 Rs).
7. Find the average of the $\bar{X}$s, or grand mean ($\bar{\bar{X}}$). This value becomes the centerline on the $\bar{X}$ chart.
8. Find the average of the ranges (R), and use it as the centerline of the R chart.
9. Select from figure 3–5 the values of A_2, D_3, and D_4, which correspond to the sample size (n), used.
10. Compute the UCL and the LCL for the charts desired using the following formulas:

$$\bar{X} \text{ chart } UCL_X = \bar{\bar{X}} + A_2\bar{R}$$

$$LCL_X = \bar{\bar{X}} - A_2\bar{R}$$

$$R \text{ chart } UCL_R = \bar{R} + D_4\bar{R}$$

$$LCL_R = \bar{R} - D_3\bar{R}$$

11. The last step is simply to plot the centerline and control limits on the chart using the vertical axis for $\bar{X}$ or R values, while the horizontal axis identifies the sample numbers (i.e., first, second, third, etc.). Out-of-control conditions are indicated in the same manner as on other control charts. As usual, when recomputing control limits, do not include items of assignable cause in the new computations.

Self-Test Questions

After you complete these questions, you may check your answers at the end of the unit.

422. Using charts for individuals

1. What is the CL for the chart for individuals?

2. When a mean ($\bar{X}$ chart) and standard deviation (s) of a sample of past data are known, how are the upper and lower control limits determined for the chart of individuals?

3. In what type of maintenance situation can the control chart for individuals be used?

4. The control limits of the chart for individuals are based on what kind of sampling distribution?

423. Using charts for averages

1. How does the centerline of a chart for individuals *compare* with the centerline of a chart for averages?

2. What is the measure of variability of a chart for averages?

3. What is the difference between the points plotted on a chart for individuals and the points plotted on a chart for averages?

4. Name one *advantage* that a chart for averages has over a chart for individuals.

5. Compare the sampling distributions associated with a chart for individuals with the sampling distribution of a chart for averages.

424. Developing charts for dispersion

1. What is the purpose of a chart for dispersion?

2. What other control chart is used in conjunction with an R chart?

3. How does subgroup or sample size affect the construction of the R chart?

4. Why must the sample size *never* be changed for previously established $\bar{X}$ chart and R charts?

3–3. Control Charts for Attributes

In this section, you will study control charts based on collection of data by the method of attributes. The data collected for study consists of only the number of items possessing, as well as the number of items not possessing, certain specified flaws—not any actual measured values. This section discusses the measurement of a selected group of samples against a control chart used as a standard. Therefore, the control charts used in the illustration represent limits using data from the same sample group. The goal of this section is to plot samples on the chart and identify out of control samples within the same group of samples.

425. Developing P charts

When discussing control charts for attributes, one needs to distinguish between the terms defect and defective. A *defect* is an error or discrepancy in an item and, therefore, detracts from its quality. An item with a defect that renders that item unserviceable is termed *defective*. Defects are not measured, but are counted. A defect either exists or does not exist, and any item may have one or more defects. For example, there may be one or more discrepancies found during an inspection. The C chart, which is discussed later, is designed to plot the number of defects per unit inspected. An item having at least one defect that renders it unserviceable is classified as defective. During inspections, items are classified as either defective or nondefective. Use the P chart to plot the *percent* of defective items. An example of a P chart in operation is shown in figure 3–6. The points plotted on the chart represent the data shown in figure 3–7. Let's assume that the data resulted from the inspection of job data documentation (JDD) work orders for errors in documentation. The four columns in figure 3–7, from left to right, show the sample number, the number of work orders inspected, which is sample size (n), the number defective or in error, and the percent defective. For example, sample number 1 contained 87 work orders. When these work orders were inspected, six were found to be defective or in error.

The quotient of $6 \div 87 = .0690$ or 6.90 percent defective, which was plotted as the first point on the P chart in figure 3–6. The percentages defective for the 24 samples listed in figure 3–7 have been plotted on the P chart in figure 3–6 in the order in which the inspections occurred.

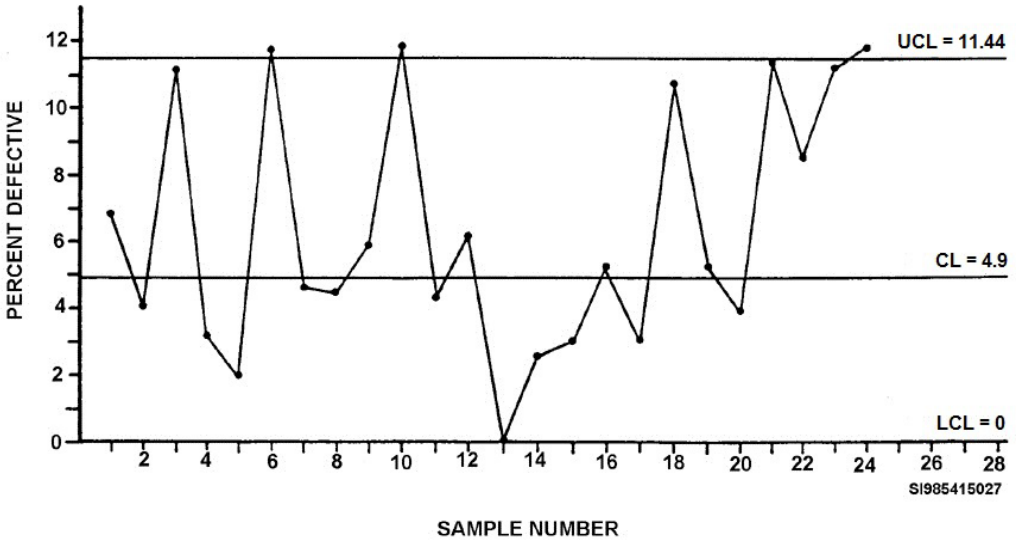


Figure 3–6. P chart.

Sample Number	Number Inspected	Number Defective	Percent Defective
1	87	6	6.9
2	98	4	4.08
3	81	9	11.11
4	97	3	3.09
5	101	2	1.98
6	111	13	11.71
7	109	5	4.59
8	89	4	4.49
9	101	6	5.94
10	85	10	11.76
11	96	4	4.17
12	82	5	6.10
13	105	0	0.00
14	118	3	2.54
15	100	3	3.00
16	95	5	5.26
17	99	3	3.03
18	93	10	10.75
19	96	5	5.2
20	102	4	3.92
21	123	14	11.38
22	94	8	8.51
23	80	9	11.25
24	110	13	11.8
TOTALS ⇒	2352	148	6.3

Figure 3–7. P chart data.

By looking at the number inspected column of figure 3–7, you can see that the number inspected varies from sample to sample. Therefore, the P chart in figure 3–6 is being run in a situation where the sample size is not constant. This can be done easily when the sample size does not vary more than about 30 percent from the average sample size (total number inspected divided by total number of samples, which in this case is 2352 ÷ 24, or 98). The centerline of the P chart is *not* affected by *changes* in sample size; *only* the control limits are affected.

If a point falls close to the approximate limit line so that you need to know the exact limit to determine whether the point is in or out of control, you can compute the exact limit for the particular sample size involved. A z-score set at three standard deviation control limits for any size sample can be computed by use of the following formulas:

$$\text{CL} = \bar{\text{P}} = \left(\frac{\text{Total number of defectives}}{\text{Total number inspected}} \right) \times 100$$

$$\text{UCL} = \bar{\text{P}} + z\sqrt{\frac{\bar{\text{P}}(100 - \bar{\text{P}})}{\text{n}}}$$

$$\text{LCL} = \bar{\text{P}} - z\sqrt{\frac{\bar{\text{P}}(100 - \bar{\text{P}})}{\text{n}}}$$

Where:

$\bar{\text{P}}$ = average percent defective or the CL value

z = 3

n = average sample size

Let's look for the CL, the UCL, and the LCL for the average sample size of 98. Figure 3–6 shows these values.

$$\text{CL} = \bar{\text{P}} = \left(\frac{148}{2352} \right) \times 100 = 6.29\% \approx 6.3\%$$

$$\text{UCL} = 6.3 + 3\sqrt{\frac{6.3(100 - 6.3)}{98}} = 13.66\%$$

$$\text{LCL} = 6.3 - 3\sqrt{\frac{6.3(100 - 6.3)}{98}} = -1.06\% \text{ or } 0\% \text{ since there is no negative percentage}$$

Since n is the denominator of the formula, an increase in the sample size causes the limits to move closer to the centerline, while a decrease in sample size causes the limits to move farther from the centerline. By keeping this relationship in mind, you can usually make a decision about a point close to the approximate limits without actually having to apply the formulas. Another factor is the number of standard deviation used. If we used a number less than 3, then our control limits narrows closer to the centerline.

Let's pick a sample number and show a computation for the UCL. Look at point number 10 in figure 3–6. That point represents a smaller sample size of 85. This tells you that the exact UCL for point 10 is farther from the centerline than the approximate UCL. Is the exact UCL far enough from the centerline to throw point 10 within limits? In this case, you cannot tell without computing the exact UCL for point 10. Make the computation by using the UCL formula as follows:

$$UCL = \overline{P} + 3\sqrt{\frac{\overline{P}(100-\overline{P})}{n}}$$

$$= 6.3 + 3\sqrt{\frac{6.3(100-6.3)}{85}}$$

$$= 6.3 + 3\sqrt{6.94}$$

$$= 6.3 + 3(2.63)$$

$$= 14.19$$

Looking at figure 3–7, note that point 10 was plotted at 11.76, which is below the exact UCL of 14.19. Therefore, point 10 is actually within control. Now look at the P chart in figure 3–6 again. The point representing sample 3 is very close to the approximate UCL. Is this point really within the limit? Note in figure 3–7 that the size of sample 3 is 81, which is smaller than 98, the sample size for which the UCL was computed. Since point 3 represents a smaller sample, you know without making any computation that the exact UCL is farther from the centerline than the approximate UCL, and point 3 is actually not as close to the exact UCL as it appears.

Now look at point 6. Is that point really out of control? Again, in figure 3–7, note that point 6 represents a sample of size 111, which is larger than 98. Therefore, the exact UCL (13.22) is closer to the centerline than the approximate UCL, but point 6 is still within limits.

Seven consecutive points on one side of the centerline gives us grounds for suspecting that the average has shifted. The points in the left half of the P chart in figure 3–6 appear to be running a little high, but you do not have seven consecutive points above the centerline. Overall, all samples are within limits using the average sample size for the centerline.

When using this type of chart, you must be aware of the anomalies that have been mentioned. If points appear out of control, check them against their actual control limits. This is also true of those points that are close to being out of control.

426. Developing C charts

The C chart is designed to plot the *number* of defects per *unit* inspected. The unit inspected can, for example, be a certain item of equipment. For this chart, the sample size or the unit inspected must remain constant. In other words, unlike the P chart, the area of opportunity for defects to occur must not change when using the C chart.

An example of a C chart in operation is shown in figure 3–8. The 12 points plotted on this chart represent the data shown in figure 3–9, which are 12 samples. These 12 samples are being measured against a C chart that was developed from previous data.

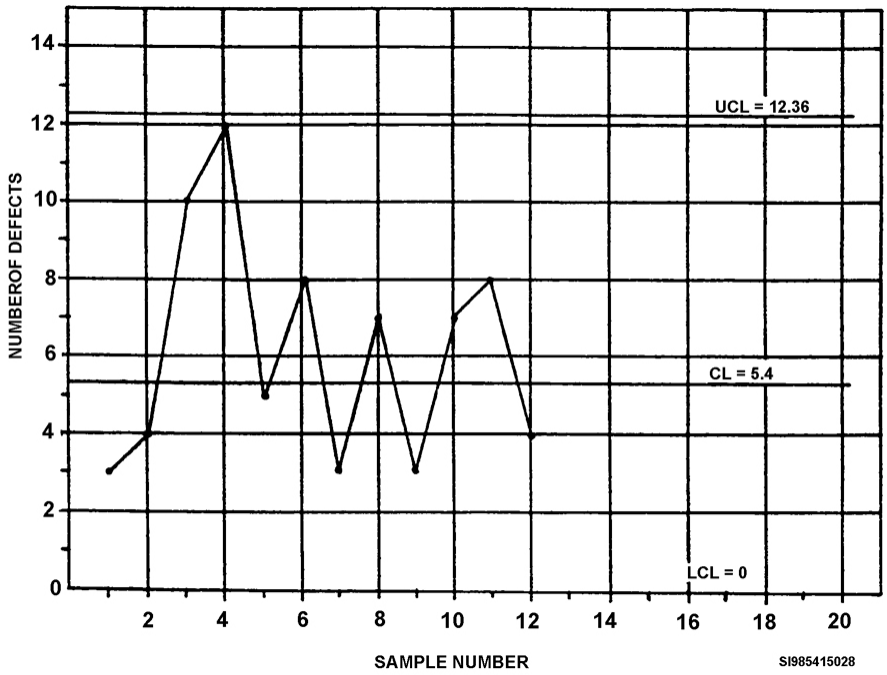


Figure 3–8. C chart.

Sample Number	Number of Defects
1	3
2	4
3	10
4	12
5	5
6	8
7	3
8	7
9	3
10	7
11	8
12	4

Figure 3–9. C chart data.

Assume the data consists of discrepancies found during a maintenance inspection. The centerline and control limits used on the C chart were computed from 20 samples of past data, and the average number of defects per unit inspected ($\bar{C}$) was used as the centerline. The following formulas are used to compute the three standard deviation control limits:

$$\bar{C} = \frac{\text{Total number of defects}}{\text{Number of units inspected}}$$

$$UCL = \bar{C} + z\sqrt{\bar{C}}$$

$$LCL = \bar{C} - z\sqrt{\bar{C}}$$

For purpose of illustration, we arbitrarily selected our centerline $\bar{C}$, the UCL, and the LCL based on the 20 hypothetical samples (or units) inspected. Control limits on the C chart serve the same purpose as the P chart. A point outside the control limits indicates the presence of assignable cause for variation.

Although a C chart and a P chart look somewhat alike, they are different in several ways. As indicated earlier, a C chart plots number of defects, while a P chart plots percent defective. The sample size or unit inspected for a C chart *must remain constant*, while the sample size for a P chart can vary. By looking at the formulas for both charts, you can see that C chart formulas are less complicated than P chart formulas.

427. Developing U charts

The U chart is an attribute-type chart primarily concerned with defects per unit when your number of inspected items varies from inspection to inspection. A defect, as previously defined, is a minor flaw or discrepancy that detracts from the quality of the item, but does not necessarily make the item inoperative. The U chart enables us to pinpoint unusual defects and thereby makes it possible to correct and possibly prevent the situation from recurring or developing into a defective process.

An example of a U chart is shown in figure 3–10. The defects per sortie are plotted on this chart represent the data shown in figure 3–11. This chart has one other unusual characteristic. The confidence interval changes for each aircraft and is dependent upon the number of sorties flown. This is because as our sampling size increases we expect less variation in the sampled group.

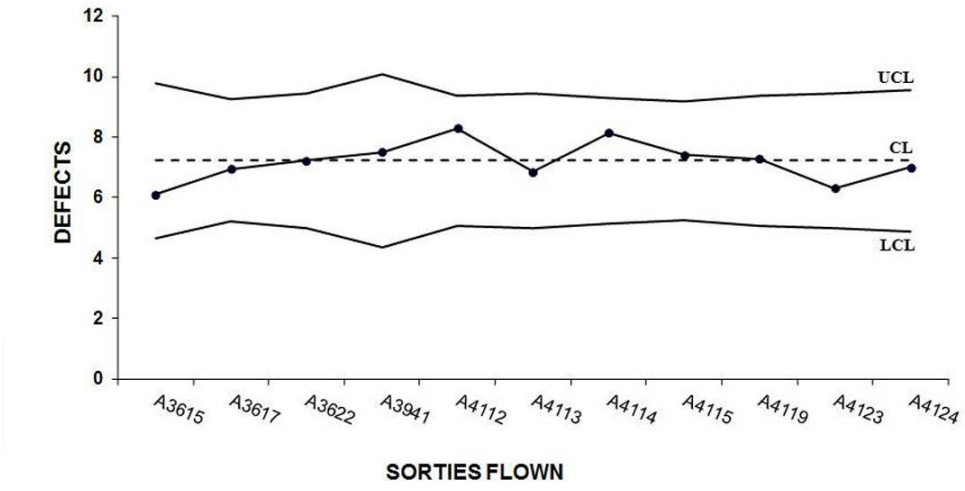


Figure 3–10. U chart.

A/C	Sorties Flown	Defects	Defects per Sortie	σU	UCL	CL	LCL
A3615	10	61	6.10	0.85	9.77	7.22	4.67
A3617	16	111	6.94	0.67	9.24	7.22	5.21
A3622	13	94	7.23	0.75	9.46	7.22	4.98
A3941	8	60	7.50	0.95	10.07	7.22	4.37
A4112	14	116	8.29	0.72	9.38	7.22	5.07
A4113	13	89	6.85	0.75	9.46	7.22	4.98
A4114	15	122	8.13	0.69	9.30	7.22	5.14
A4115	17	126	7.41	0.65	9.18	7.22	5.27
A4119	14	102	7.29	0.72	9.38	7.22	5.07
A4123	13	82	6.31	0.75	9.46	7.22	4.98
A4124	12	84	7.00	0.78	9.55	7.22	4.89
Z	TOTAL	TOTAL					
3	145	1047					

Figure 3–11. U chart data.

Although the U chart is not used extensively, it still has some valid uses when you need to calculate defects per sortie. This is a variation of the C chart. Use the following formulas when building the U chart.

$$U = \frac{Number\ of\ Defects}{Number\ of\ Sorties}$$

$$\bar{U} = \frac{Total\ Number\ of\ Defects}{Total\ Number\ of\ Sorties}$$

$$UCL = \bar{U} + (z)(\sigma U)$$

$$LCL = \bar{U} - (z)(\sigma U)$$

For your upper and lower control, you can use a z of 3 in most cases. If you do not use z=3, you will need to choose what your z will be depending on your desired confidence interval.

428. Maintenance applications for control charts

Finding maintenance applications for control charts is not an easy task. It requires considerable control chart knowledge plus the use of your imagination and common sense. Two factors that might be considered are (1) the nature of the maintenance process under analysis and (2) the purpose of the control chart itself. One of your basic guidelines in making control chart applications is to have the event repeated often and have the circumstances surrounding each repetition of the event as much alike as possible. In maintenance applications, as you know, a compromise is usually necessary. First, let's summarize the control charts used in this lesson and examine situations in which these charts can be applied.

Charts for variables

The three charts for variables that you studied were a chart for individuals, a chart for averages, and a chart for dispersion. All of these control charts can be used to monitor repetitive tasks and inspections where measurements are made in man-hours or clock hours. Individual measurements are plotted on a chart of individuals. This chart can be used in situations where there is hardly enough data for a chart

for averages. Averages of small samples are plotted on a chart for averages. It works well when there is a relatively large amount of data. The range chart controls the dispersion of a process and is *normally* used with a chart for averages.

Charts for attributes

Three control charts for attributes were discussed—the P chart, the C chart, and the U chart. You use the P chart in situations where the inspection involves classifying items as defective or nondefective and showing that classification as a percent. Sample size may vary and the percent defective is plotted. The C chart is used in situations where inspection involves counting the number of defects or discrepancies per unit inspected. The unit inspected or sample size must remain constant. The *number* of defects per *unit* inspected is plotted on a C chart. The U chart is used when the number or size of your inspection items change and the inspection involves classifying items as defective or nondefective. The U chart is an attribute type chart primarily concerned with defects per unit when your number of inspected items varies from inspection to inspection.

Self-Test Questions

After you complete these questions, you may check your answers at the end of the unit.

425. Developing P charts

1. Define the terms *defect* and *defective*.

2. Using the information shown in figure 3–7, answer the following questions concerning interpretation of points in the right half of figure 3–6. Remember that the UCL in figure 3–6 was computed for an average sample size of 98. Use the UCL formula only when necessary:

   a. Is point 21 in control? Give a reason for your conclusion.

   b. Is point 23 actually within limits? Explain why or why not.

   c. Is point 24 actually out of control? Explain why or why not.

426. Developing C charts

1. What is the C chart designed for?

2. What does the C chart use for its centerline?

3. What condition *must* the sample size or unit inspected meet when using the C chart?

427. Developing U charts

1. What is the U chart's primary concern when looking at defects per unit?

2. What unusual characteristic does the U chart have?

3. The U chart is a variation of what control chart?

428. Maintenance applications for control charts

1. For the following situations, indicate which type of control chart is *best* suited:

   a. Man-hours required to perform an inspection are being collected. You want to plot each man-hour value because there is not very much data available.

   b. You are collecting small samples of data of a repetitive job from a wide range of data and you want to plot a representative value.

   c. Samples of AFTO forms are being inspected and classified as either correct or incorrect.

   d. The data being collected consist of a number of discrepancies per unit inspected.

Answers to Self-Test Questions

420

1. It involves a quantitative classification of data.
2. It is a nonquantitative classification.
3. Small influences, considered a natural part of the maintenance process.
4. Specific disorders within a maintenance process that can be identified and possibly eliminated.
5. To detect the presence of assignable causes for variation in a maintenance process.

421

1. Using guidelines and setting control limits
2. To determine if the point was due to chance or assignable cause.
3. Too much time may be spent looking for problems that do not exist.
4. Switch to tighter controls.

422

1. The mean.
2. By adding and subtracting one or more standard deviations to and from the mean.

3. In repetitive-type jobs or where the distribution of the individual values is fairly symmetrical.
4. Normal distribution.

423

1. The centerlines are the same.
2. Standard error of the mean.
3. Plots on a chart for averages represent sample mean values, while plots on a chart for individuals represent individual values.
4. The chart for averages is less likely to be distorted by a skewed population.
5. Both are based on the normal distribution.

424

1. To measure changes in the dispersion of a process.
2. $\bar{X}$ chart.
3. Subgroup size should be maintained relatively small to keep variations within subgroups to a minimum.
4. Control limits are based on sample size.

425

1. A defect is a discrepancy that detracts from the quality of an item. A defective item is an item with a defect that renders it unserviceable.
2. a. Yes. The exact UCL is 12.87.
   b. Yes. Because of the smaller sample, the exact UCL is farther from the mean than with sample size of 98.
   c. No, the exact UCL is 13.25.

426

1. To plot the number of defects per unit inspected.
2. The average number of defects per unit.
3. They must remain constant.

427

1. When your number of inspected items varies from inspection to inspection.
2. The confidence interval changes due to an increase in sampling size.
3. The C chart.

428

1. a. Chart for individuals.
   b. Chart for averages.
   c. P chart.
   d. C chart.

Complete the unit review exercises before going to the next unit.

Unit Review Exercises

Note to Student: Consider all choices carefully, select the *best* answer to each question, and *circle* the corresponding letter. When you have completed all unit review exercises, transfer your answers to the Field-Scoring Answer Sheet.

Do not return your answer sheet to AFCDA.

52. (420) In general statistical terms, a control chart tells you
    a. what the problem is.
    b. when to look for a problem.
    c. where the root cause of a problem is.
    d. how to correct and eliminate a problem.

53. (421) What normally happens when identifying processes out of control and the control limits are set too close together?
    a. The analyst fails to look for problems that exist.
    b. Nothing, since processes are rarely out of control.
    c. Too much time is spent adjusting the control limits.
    d. Too much time is spent looking for problems that do not exist.

54. (421) When identifying processes out of control and using a control chart, what action should you take if you have set your control limits at *three standard deviations* and later find not enough time is spent looking for assignable causes?
    a. Switch to tighter limits.
    b. Recalculate the standard deviation.
    c. Keep established limits and do not investigate.
    d. Discard current data and recalculate the standard deviation.

55. (422) In statistical terms, what does the control chart for plotting individual X values *use* for the centerline?
    a. Mean.
    b. Mode.
    c. Standard deviation.
    d. Standard error of the mean.

56. (422) The statistical interpretation of a control chart for individuals would be *distorted* if the
    a. distribution is normal.
    b. distribution is extremely skewed.
    c. standard deviation is too small.
    d. standard deviation is too large.

57. (423) In statistics, the control chart for averages is *best* used when dealing with data that is
    a. simple.
    b. complex.
    c. in relatively small amounts.
    d. in relatively large amounts.

58. (424) Which statistical chart measures changes in the means of a series?
    a. Chart for individuals.
    b. Chart for averages.
    c. Range chart.
    d. P chart.

59. (424) What subgroup sample size is *commonly* used when constructing a statistical range chart?
    a. 4 or 5.
    b. 6 or 10.
    c. 10 or 20.
    d. 20 or 25.

60. (425) What statistical control chart is used to plot the *percent* of defective items?
    a. C chart.
    b. P chart.
    c. Chart of averages.
    d. Chart of individuals.

61. (425) On a P chart in statistics, changes in sample size are affected by the
    a. centerline.
    b. control limits.
    c. standard deviation.
    d. standard error of the mean.

62. (426) In statistics, a C chart may be used to measure the
    a. number of defective units.
    b. number of defects per unit.
    c. percent of defective units.
    d. percent of defects per unit.

63. (426) In statistics, one difference between a C chart and a P chart is the C chart
    a. is more accurate.
    b. formulas are more complicated.
    c. sample size must remain constant.
    d. measures variables while the P chart measures attributes.

64. (427) What unusual characteristic does the U chart have?
    a. The confidence interval does not change due to an increase in sample size.
    b. The confidence interval changes due to an increase in sampling size.
    c. The upper and lower control limits are fixed.
    d. The z score increases due to sampling size.

65. (428) In statistics, a range chart is *normally* used with a
    a. C chart.
    b. P chart.
    c. chart for averages.
    d. chart for individuals.

66. (428) In statistics, what are the three control charts for attributes?
    a. C, R, and X charts.
    b. C, P, and U charts.
    c. U, R, and P charts.
    d. X, C, and P charts.

67. (428) On what statistical control chart is the *number* of defects per *unit* plotted?
    a. C chart.
    b. P chart.
    c. Chart for averages.
    d. Chart for individuals.

Please read the unit menu for unit 4 and continue ➔

Unit 4. Predictive Analysis

AN IMPORTANT PART of statistical analysis is to make valid, informed predictions based on past data and precise methods of statistical prediction. This unit begins with correlation, which is used to establish a relationship between two sets of data. Although some degree of relationship will always exist, your concern is the amount of correlation and its significance. Next, you will plot data to display an increasing, decreasing, or continuing stable trend. Using trend analysis techniques, you will be able to draw a trend line to reflect the direction data is taking and estimate future occurrences. Then by using the extrapolation technique, you will be able to expand a trend out into the future. Another extremely useful tool is called regression analysis, this helps in predicting future trends. As a final topic, we will take a look at the science of probability which is predicting the likelihood that something will occur.

4–1. Correlation Analysis

In this section, you will learn how correlation is used as a statistical measurement of the relationship between two data sets. You will study correlation concepts and two methods of correlation analysis: Pearson's product-moment correlation and Spearman's rank order correlation. You will attempt to determine if and how two sets of data are related and to what extent.

429. Correlation concepts

Correlation is a statistical technique that measures whether there is a relationship between two or more sets of data and, if so, how strong is the relationship. For example, intelligence quotient (IQ) and grades achieved in course work are related. You use this relationship in two ways. First, you would expect that the higher a person's IQ, the higher should be his or her grades. Second, if you

happen to know a person's IQ, then you can predict what grade he or she might be expected to get in a course.

Other examples of relationships are amount of training versus length of time to complete training, gas mileage versus speed at which traveled, equipment failures versus operating hours, and aircraft speed versus flight time. As you can readily see, each member of each pair of data is related to the other member of the pair. You would normally expect that the more training people need, the more time it would take. Or, you would expect that the faster a car travels, the lower the gas mileage (fewer miles per gallon) should be. Similarly, the more hours an item operates, the more failures should occur, or the faster an aircraft flies, the less time it should take to arrive at a destination.

You can use correlation to show the degree of relationship or association between two sets of measurements. The two sets of measurements make up two frequency distributions, each of which has its own mean, median, mode, standard deviation, and so on. The groups upon which the measurements are made may consist of people, pieces of equipment, periods of time, and so forth. Before you can compute correlation, you must have available a pair of measures on each member of the group. For example, if you are interested in computing the correlation between not mission capable supply (NMCS) rates and cannibalizations to see if, in fact, high cannibalizations result from high NMCS rates, you need a pair of measures (cannibalizations and NMCS rates) for the time period under study.

Finding a correlation between the two sets of data does not prove that either one caused the other. A correlation between two sets of data may suggest the possibility that one causes the other, but no more. The relationship may even be accidental. The proof of what caused the two sets of data to be related must be obtained by other means, such as a study or a referral.

Limitations and types of correlation

Although there is a relationship between both elements of a pair, you cannot conclude that one caused the other. For example, we mentioned the correlation between a person's success and the amount of training. Even though it might seem logical to say the more training a person has, the more successful he or she should be; we cannot say that training causes a person to be successful. Other things help cause success besides training—things such as luck, hard work, or being in the right place at the right time. Similarly, although it seems logical that the faster a person drives a car, the lower the gas mileage will be, we cannot say that speed causes the decrease in gas mileage. Other factors, such as timing, type of gas, dirty plugs, and so forth, help cause a decrease in gas mileage. It may be that one factor causes another or that some other factor causes both, but you must obtain proof of this causation by means other than correlation. Correlation can only suggest causal relationship and how much of a relationship exists, but no more.

Another limitation is that the elements you are trying to relate must have something in common and make sense. For example, a high degree of correlation might exist between the amount of whiskey produced in Scotland and the divorce rate of a certain class of people, but these two series of data have nothing in common. It just so happens that the two series of data both rise during a certain period each year. Two other elements or series of data that would make no sense to correlate would be a person's height and IQ. When you think about this relationship, you realize that there is very little possibility that these two factors are biologically related. Ensure your data are related before applying correlation techniques. There are two types of correlation —*linear* and *curvilinear*.

Linear correlation

Linear correlation is the degree of relationship between two variables that have a constant variation. This linear correlation is either positive (direct) or negative (indirect). *Positive correlation* is a condition where, as one variable increases, the other variable also increases. Likewise, as one variable decreases, so will the other. One such maintenance example is, the more reliable the maintenance, the higher the system capability. On the other hand, *negative correlation* results when two variables

move in opposite directions at the same rate. In other words, as one variable increases the other decreases and vice versa. For example, an increase in the number of inspection man-hours could cause a decrease in failures on a particular end item.

Curvilinear correlation

Curvilinear correlation is the relationship or association between two or more variables or attributes that tend to be nonlinear. For example, as one variable increases, the other variable could increase up to a point and then begin to decrease, as the first variable continues to increase. Or, the rate at which the second variable changes may change (i.e., get larger or smaller while the first variable continues to change at a constant rate). Probably one of the most common examples of this type of correlation is that of vehicle speed versus fuel consumption.

Coefficient of correlation

The coefficient of correlation is a number that tells to what extent two sets of data are related. The correlation coefficient can fall anywhere between negative one and positive one ($-1 < 0 < +1$). Since the number can vary from +1.00 through 0 to –1.00, it can be a decimal value, such as +.92, –.45, –.85, and so forth. The closer this decimal value is to +1.00 or –1.00, the higher the relationship or correlation. The *closer* it approaches zero, such as +.25 or –.15, the *lower* the correlation, which usually indicates *very little relationship*. The sign of the value indicates whether the relationship is positive or negative. A negative value represents a decreasing trend, and a positive value represents an increasing trend. The positive and negative signs depict the direction of the trend—not the degree of relationship.

Positive relationship

A positive relationship means that as values in one set of measures *increase*, the corresponding or paired values in the other set *also increase*. In other words, the high values in one set tend to be paired with the high values in the other set, and low values tend to be paired with low values. For example, high cannibalizations can be positively related to high NMCS rates.

Negative relationship

A negative relationship means that as values in one set of measures increase, the corresponding or paired values in the other set decrease. In other words, the high values in one set tend to be paired with low values in the other set, and low values tend to be paired with high values. For example, suppose for certain equipment the mean time between failures is negatively related to the age of the equipment. If this is true, the greater the age of the equipment, the less the average time between failures.

Scatter diagram

The easiest approach to the study of the correlation between two sets of data is to make a simple scatter diagram—a graph or picture showing the relationship *between* two sets of measures. Let's examine its construction and interpretation. For example, we developed a table of related data (fig. 4–1). Then we developed a scatter diagram (fig. 4–2) constructed from the related data in figure 4–1. Each point on this scatter diagram represents one pair of X and Y values. The point representing the first pair of X and Y values (2,2) is plotted in the following way: draw a vertical line up from the X value of 2 located on the X axis and a horizontal line to the right from the Y value of 2 located on the Y axis. The point representing the X value of 2 and the Y value of 2 is plotted where these two lines meet (fig. 4–2).

RELATED DATA

X	Y
2	2
4	8
5	6
5	12
6	10
7	10
7	14
7	16
9	16
10	20

Figure 4–1. Related data.

Construction

The point representing the second pair of X and Y values (4, 8) is plotted in the same way. You draw a vertical line up from the X value (4) located on the X axis and a horizontal line to the right from the Y value (8) located on the Y axis. The point representing the X value of 4 and Y value of 8 is plotted where these two lines meet (fig. 4–2). The points representing the remaining pairs of X and Y values are plotted in a similar manner. From this scatter diagram, you get a better idea of the relationship between the two sets of data than you get by just looking at the pairs of X and Y values in figure 4–1.

Interpretation

The scatter diagram in figure 4–2 illustrates several things about the relationship that otherwise would not be readily seen. First, high X values tend to be associated with high Y values, and low X values tend to be associated with low Y values. Second, this relationship, although reasonably high, is not perfect. The coefficient of correlation representing this relationship would probably be close to +.90. Third, you can describe the general trend of this relationship by drawing a straight line from lower left to upper right. When the plotted points tend to follow a straight line from lower left to upper right, the relationship is positive and linear.

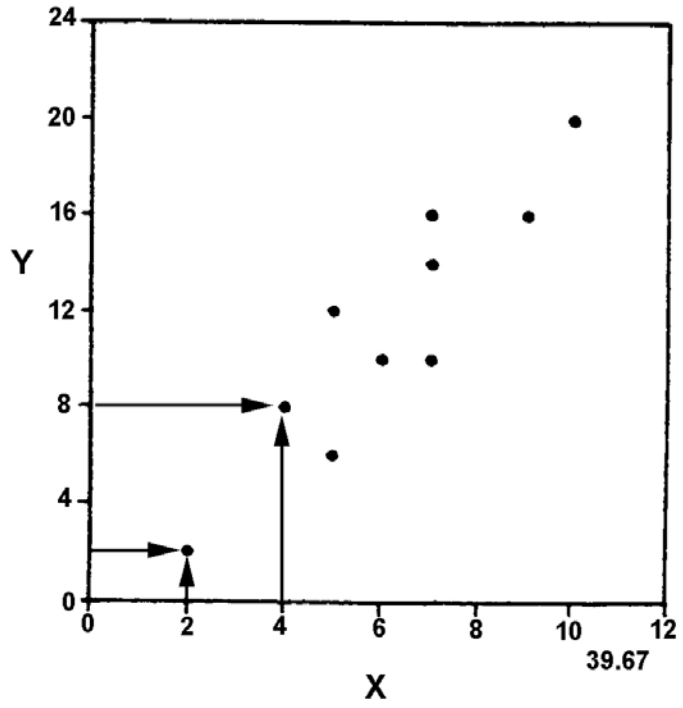


Figure 4–2. Scatter diagram construction.

On the other hand, when the plotted points tend to follow a straight line from upper left to lower right (fig. 4–3), the relationship is negative and linear. Here again, the relationship is reasonably high but not perfect. High X values tend to be associated with low Y values, and low X values tend to be associated with high Y values. The coefficient of correlation representing this relationship is probably close to –.90. Notice that the width of the scatter of the points in figure 4–2 is about the same as the width of the scatter of the points in figure 4–3. The scatter diagrams in both figures show about the same amount of relationship. In figure 4–2, the relationship is positive, but in figure 4–3, it is negative.

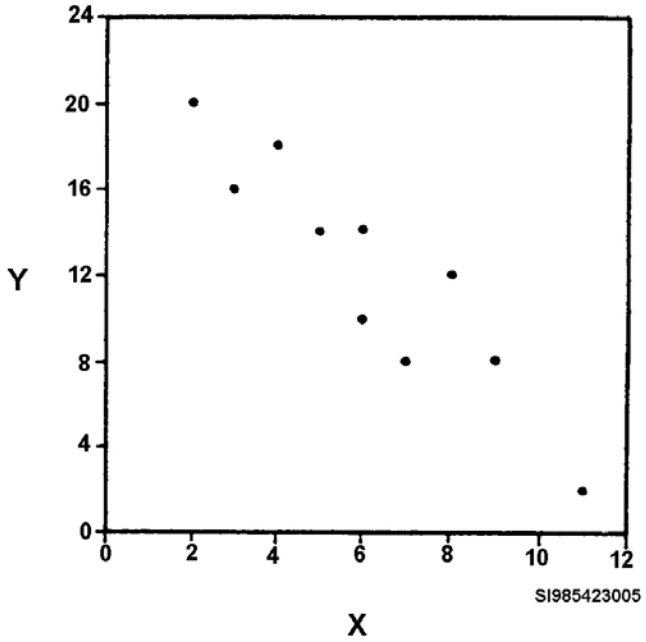


Figure 4–3. Scatter diagram (negative relationship).

When the plotted points fall exactly on a straight line extending from lower left to upper right (fig. 4–4), the relationship is perfect positive. The correlation coefficient for a perfect positive relationship is a +1.00. Perfect positive and perfect negative correlations are the highest correlations possible.

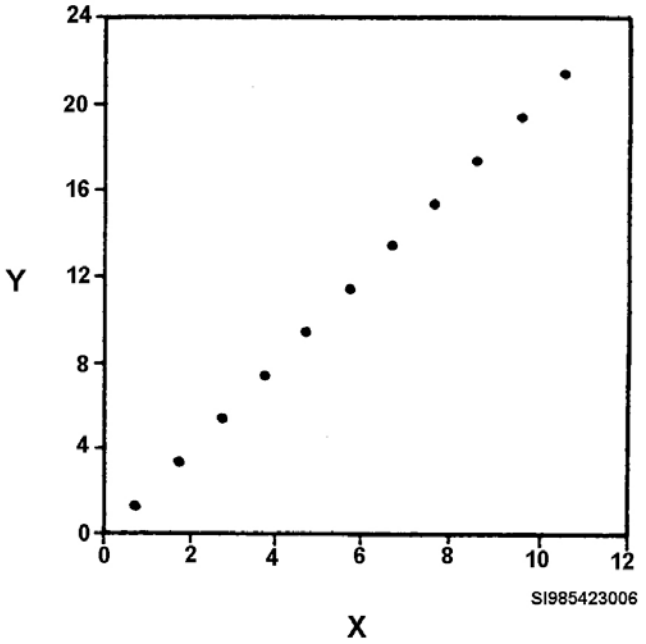


Figure 4–4. Scatter diagram (perfect positive relationship).

A perfect negative relationship is represented in figure 4–5. Notice that all the points are on a straight line extending from upper left to lower right, depicting a perfect negative correlation. The coefficient of correlation for a perfect negative correlation is –1.00, meaning that the highest X value is associated with the lowest Y value, the second highest X value with the second lowest Y value, the third highest X value with the third lowest Y value, and so on. Perfect positive and perfect negative relationships are rare.

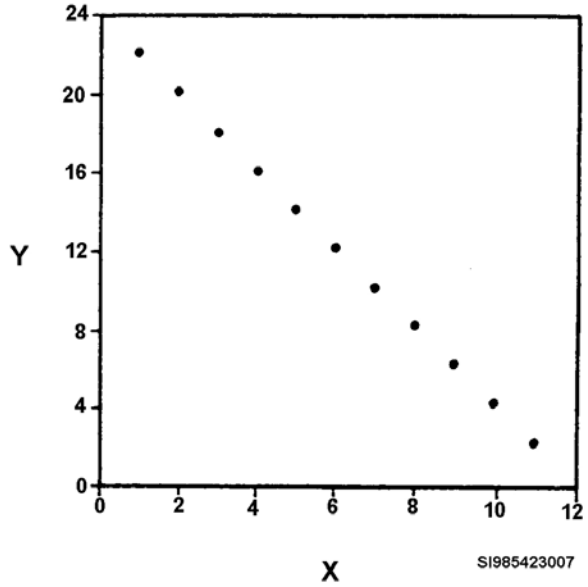


Figure 4–5. Scatter diagram (perfect negative relationship).

At the other extreme from a perfect relationship (either positive or negative) is the zero correlation. When the points on a scatter diagram are spread evenly in all directions (fig. 4–6), you have a case of little or no relationship. The coefficient of correlation representing this situation would be near .00, indicating that a given X value is probably not associated with any Y value. In other words, the scatter diagram shows no evidence of either a positive or a negative relationship.

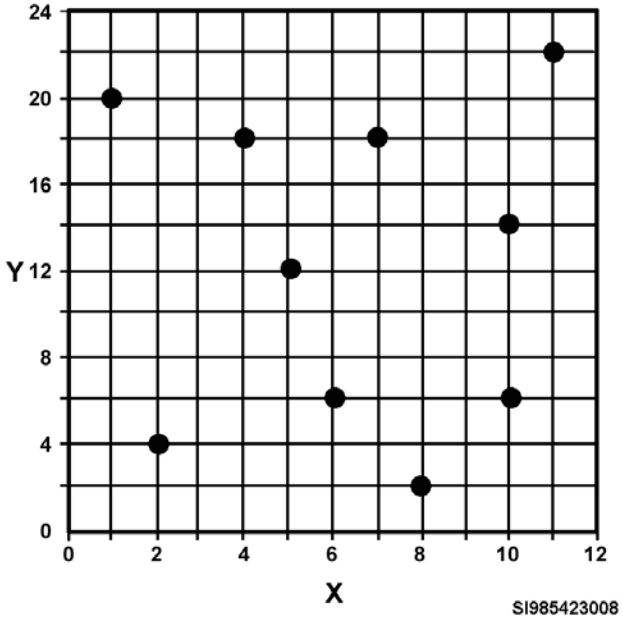


Figure 4–6. Scatter diagram (little/no relationship).

The relationships discussed so far have been linear. There are situations where the trend is best described by a curved line, making the relationship *nonlinear* or *curvilinear*. An example of a nonlinear relationship is shown in figure 4–7. Notice that as X increases, Y decreases rapidly at first. But then, as X continues to increase, Y decreases at a slower and slower rate until further increases in X are accompanied by almost no change in Y. In the case of a nonlinear curve displayed by a scatter

diagram, the coefficient of correlation is not effective because it applies only to linear curves. A nonlinear curve does not necessarily mean that there is no relationship between the two data sets being measured. Other methods of trend analysis are used.

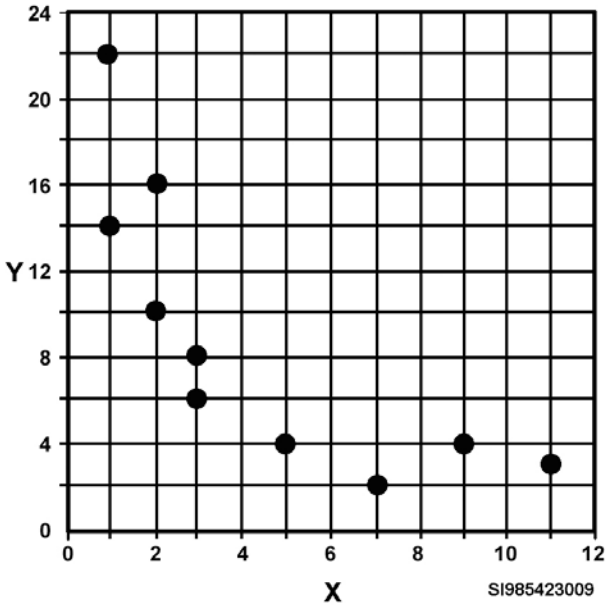


Figure 4–7. Scatter diagram (nonlinear).

By making a simple scatter diagram, you can determine several things. First, you can decide whether there is an apparent relationship. Second, you can see whether a relationship is linear or nonlinear. Third, you can determine whether the correlation is positive or negative. Fourth, you can decide whether it is high or low.

After studying the scatter diagram, you may get ideas for further analysis. With a little knowledge of correlation and some practice, you should be able to make good use of the scatter diagram.

430. Performing Pearson's product-moment correlation

Previously, we used a scatter diagram to find out if there is a strong relationship between two data sets. We use the scatter diagram as an initial check or a rough idea of where the correlation is heading.

When there is a strong indication of a linear relationship, such that a line going in one direction, then you can measure the strength of the correlation with the use of Pearson's product-moment correlation. This will give you the exact value of the correlation, which is referred to as the *coefficient of correlation.*

Pearson's coefficient of correlation is symbolized by the letter "r" and is used to measure the degree of association or linear relationship between two variables. Pearson's coefficient of correlation is only associated with linear correlation and continuous data. Pearson's method uses the actual, individual values of X and Y and can be used with samples or the population. When using samples, they must be large enough to represent the population. Samples must also meet the following criteria before Pearson's product-moment correlation techniques can be applied:

1. The data must have been selected at random and in pairs from a normal distribution.
2. The data must display homogeneity of variance.
3. It can only come from an interval or ratio measurement scale.

When measuring data relationships, you generally denote the Y quantity as the dependent variable and the X quantity as the independent variable. Regardless of which variable is dependent, the procedure for determining the coefficient of correlation is the same. Use the following formula to compute Pearson's r:

$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}}$$

Where:

r = coefficient of correlation

N = number of pairs of data

X = independent values

Y = dependent values

Use this formula with the data in figure 4–8 to calculate r. The data represents a component operating time by month (X values), compared to its mean time between failures (MTBF) or Y values.

Pair No.	X	Y	X^2	Y^2	XY
1	4	7.0	16	49.00	28.0
2	6	8.0	36	64.00	48.0
3	7	9.5	49	90.25	66.5
4	12	11.0	144	121.00	132.0
5	19	13.5	361	182.25	256.5
6	23	16.5	529	272.25	379.5
7	26	20.0	676	400.00	520.0
8	32	21.0	1024	441.00	672.0
9	37	22.5	1369	506.25	832.5
10	39	24.5	1521	600.25	955.5
	$\Sigma X = 205$	$\Sigma Y = 153.5$	$\Sigma X^2 = 5725$	$\Sigma Y^2 = 2726.25$	$\Sigma XY = 3890.5$

Figure 4–8. Data method for Pearson's correlation.

The easiest way to compute the coefficient of correlation is to set up the necessary data in columns. From the formula for r, you can see that the values of X, Y, X^2, Y^2, and XY are required. Therefore, set up the necessary columns, as shown in figure 4–8. Square each X value to obtain the values for the X^2 column. For example, the first X value is 4. Square 4 and enter the results (16) in the X^2 column. Similarly, square each Y value to obtain the values for the Y^2 column. The first Y value is 7.0. Enter the square of 7.0 (49.00) in the Y^2 column. The last column is headed XY. Obtain the values for this column by multiplying the X value of each pair by the respective Y value of that pair. For example, the first X value is 4 and the first Y value is 7. Multiply these two values together and record the answer (28.0) in the XY column. Now, complete the calculation for the remaining nine pairs of data. Add the values in X, Y, X^2, Y^2, and XY columns to get ΣX, ΣY, ΣX^2, ΣY^2, and ΣXY for the formula.

From figure 4–8, the sums of the values are:

$$N = 10$$
$$\Sigma X = 205$$
$$\Sigma Y = 153.5$$
$$\Sigma X^2 = 5725$$
$$\Sigma Y^2 = 2726.25$$
$$\Sigma XY = 3890.5$$

Having found the necessary sums, substitute these values into the formula and solve for Pearson's coefficient of correlation:

$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}}$$

$$r = \frac{10(3890.5) - (205)(153.5)}{\sqrt{[10(5725) - (205)^2][10(2726.25) - (153.5)^2]}}$$

$$r = \frac{38905 - 31467.5}{\sqrt{(57250 - 42025)(27262.5 - 23562.25)}}$$

$$r = \frac{7437.5}{\sqrt{(15225)(3700.25)}}$$

$$r = \frac{7437.5}{\sqrt{56336306.25}}$$

$$r = \frac{7437.5}{7505.75}$$

$$r = 0.9909 \approx 0.99$$

The value of r is the measure of the distance a set of points varies from a straight line. Knowing the value of r, you can determine the probability of a significant relationship between the X and Y values. Figure 4–9 shows the correlation probabilities for various degrees of freedom (df) and levels of significance. Compute df for the table by subtracting 2 from the total number of data pairs:

$$df = n - 2.$$

In the example, the value of Pearson's r was .99 for 10 pairs of data. According to figure 4–9, the *minimum* value of r is 0.7646, with a df of 8 (n – 2 = 10 – 2 = 8 pairs of data), at the 0.01 level. The calculated value of .99 was much *higher* than the *minimum* required value (0.7646). As a result, you know that there is a probability of *greater* than 99 percent that a significant relationship *exists* between the X and Y values.

The coefficient of correlation by itself is not enough information to judge if a significant relationship exists between two series of data. In addition to the value of r, the number of pairs of data (sample size) is also important. For example, suppose you had a coefficient of correlation of 0.75. If three pairs of data were used to obtain this coefficient of correlation rather than 10, do you suppose the smaller sample would be as accurate as the larger? Actually, it would not. The fewer the pairs of data you have, the higher the coefficient of correlation must be before a significant relationship can exist.

Conversely, the more pairs of data you have, the lower the coefficient of correlation can be and still have a significant relationship. The relationship between the number of pairs of data (as represented by df) and the size of coefficient of correlation is shown figure 4–9.

If there is a significant relationship, then you perform regression analysis. The correlation techniques involve plotting a line of regression to show the trend of the data. Before looking at the line of regression calculation, let's briefly review correlation procedures.

df	Levels of Significance				
	.1	.05	.02	.01	.001
1	.9876	.9969	.9995	.9998	.9999
2	.9000	.9500	.9800	.9900	.9990
3	.8054	.8783	.9343	.9587	.9911
4	.7293	.8114	.8822	.9172	.9740
5	.6694	.7545	.8329	.8745	.9507
6	.6215	.7067	.7887	.8343	.9249
7	.5822	.6664	.7498	.7977	.8982
8	.5494	.6319	.7155	.7646	.8721
9	.5214	.6021	.6851	.7348	.8471
10	.4973	.5760	.6581	.7079	.8233
11	.4762	.5529	.6339	.6835	.8010
12	.4575	.5324	.6120	.6614	.7800
13	.4409	.5139	.5923	.6411	.7603
14	.4259	.4973	.5742	.6226	.7420
15	.4124	.4821	.5577	.6055	.7246
16	.4000	.4683	.5425	.5897	.7084
17	.3887	.4555	.5285	.5751	.6932
18	.3783	.4438	.5155	.5614	.6787
19	.3687	.4329	.5034	.5487	.6652
20	.3598	.4227	.4921	.5368	.6524
25	.3233	.3809	.4451	.4869	.5974
30	.2960	.3494	.4093	.4487	.5541
35	.2746	.3246	.3810	.4182	.5189
40	.2573	.3044	.3578	.3932	.4896
45	.2428	.2875	.3384	.3721	.4648
50	.2306	.2732	.3218	.3541	.4433
60	.2108	.2500	.2948	.3248	.4078
70	.1954	.2319	.2737	.3017	.3799
80	.1829	2172	.2565	.2830	.3568
90	.1726	.2050	.2422	.2673	.3375
100	.1638	.1946	.2301	.2540	.3211

Figure 4–9. Table of Critical Values for Pearson's "r."

First, examine the two series of data to determine if they have something in common and if it makes sense to correlate them. Next, determine which series is independent and which is dependent. Then, calculate the coefficient of correlation and compare it to the minimum value of r from the table of Critical Values for Pearson's r (fig. 4–9; for the appropriate n – 2). If the calculated value of r exceeds the table value, you can be confident that a significant relationship exists.

431. Performing Spearman's rank correlation coefficient

The Spearman method is an easy way to measure the degree of relationship between related sets of data. It is well suited for situations where the number of pairs of data is 30 or less. This method can be applied to data that does not show normality, is non-linear and from at least the ordinal measurement scale. Spearman's method de-emphasizes extreme values when the data is ranked and used subsequently as an estimate of the degree of correlation. Its easy calculation makes it an excellent tool for determining if correlation exists between data sets.

Computation

The correlation coefficient computed by Spearman's method is called rho (ρ). The computed value of rho is normally smaller than the computed value of Pearson's r (coefficient of correlation using the Pearson's product-moment method).

Spearman's method is *not* applicable in all situations. Since, Spearman's correlation method involves ranking and uses small samples and the ordinal measurement scale, do *not* use it as a *substitute* for the coefficient of correlation computed by the product-moment method (Pearson's) if the result is to be used in *additional* statistical testing. Spearman's method has several disadvantages:

1. It loses its usefulness when used with large samples.
2. Its calculation becomes laborious with large samples.
3. It distorts the coefficient if too many ties exist.

Let's look at the formula for Spearman's correlation. Suppose you want to know the relationship between the hours of operation and the number of failures on an end item. Let's use the data in figure 4–10 to illustrate the computation of Spearman's rank correlation coefficient. Columns (2) and (3) show hours of operation and number of failures, respectively, for the months listed in column (1).

(1)	(2)	(3)	(4)	(5)	(6)	(7)
Month	X Hours Operation	Y Number Failures	R_X	R_Y	D	D^2
Jan	460	5	1	1	0	0
Feb	520	8	2	3	1	1
Mar	560	6	3	2	1	1
Apr	840	14	5	6.5	1.5	2.25
May	750	12	4	5	1	1
Jun	960	17	8	8	0	0
Jul	920	14	7	6.5	0.5	0.25
Aug	890	10	6	4	2	4
						$\Sigma D^2 = 9.50$

Figure 4–10. Spearman's correlation table.

To compute Spearman's coefficient, go through the following steps:

1. Rank the X values (hours of operation) in column 2 and place the ranks in column 4. Assign rank one to the lowest X value (460), rank two to the second lowest X value (520), rank three to the third lowest value (560), etc. Continue until all the X values are ranked as shown in column 4 headed R_X (symbol for rank of X).
2. Rank the Y values (number of failures) in column 3 and place the ranks in column 5 heading by R_Y (rank of Y). Again, assign rank one to the lowest Y value (5), rank two to the second lowest value (6), and so on. Notice, however, in column 3 that there are two Y values that are equal, 14 and 14. Give these two values the average ranks or positions they occupy; the ranks 6 and 7. Average the two ranks (6 and 7), and assign the value 6.5 to both Y values as shown in column 5.
3. Determine the difference (D) between each pair of ranks and list the differences in column 6. The signs (+, –) are unimportant because you square the differences later.
4. Square each of the differences listed in column 6 and list the squares of the differences (D^2) in column 7. Next, add all the values in column 7 to find the sum of D^2.
5. Use the following formula to compute Spearman's rank correlation coefficient:

$$\rho = 1 - \frac{6(\Sigma D^2)}{N(N^2 - 1)}$$

Where:

N= the number of pairs.

6 = a constant.

Substitute the required values and proceed as follows:

$$\rho = 1 - \frac{6(9.5)}{\mathbf{8}(8^2 - 1)}$$

$$\rho = 1 - \frac{57}{\mathbf{8}(63)}$$

$$\rho = 1 - \frac{57}{504}$$

$$\rho = 1 - 0.11$$

$$\rho = 1 - 0.11$$

$$\rho = 0.89$$

Interpretation

What does a correlation coefficient of 0.89 mean? It means that you have a certain degree of relationship between the hours of operation and the number of failures. Interpret Spearman's rank correlation coefficient in the same way as Pearson's coefficient. Rho *varies* from –1 through 0 to +1 and is *usually* smaller than Pearson's r. Since rho is a positive number (0.89), the relationship is positive. It means that as one factor increases or decreases, you can expect the other to do the same to a certain degree. If it is a +1 correlation coefficient, the pairs of data would move up and down in direct proportion to each other. With a 0.89 positive coefficient, the same movement would still prevail but not always in direct proportion.

Testing the correlation coefficient

You know there is a relationship, but is it significant enough to warrant further tests and possibly an investigation? Use the sampling distribution in the following table of "critical values of Spearman's rank correlation coefficient." For study purposes, we show up to 24 pairs only. Remember, you can use Spearman's correlation coefficient with 30 pairs or less.

Critical Values of Spearman's Rank Correlation Coefficient		
Number of Pairs	Alpha	
N	.05	.01
4	1.000	-
5	.900	1.000
6	.829	.943
7	.714	.893
8	.643	.833
9	.600	.783
10	.564	.746
12	.506	.712
14	.456	.645
16	.425	.601
18	.399	.564
20	.377	.534
22	.359	.508
24	.343	.485

Compare the computed ρ to the critical value based on the level of significance (alpha) that we set and the number of pairs in the data. If the computed ρ > critical value then you know that there is a significant correlation between the two sets of data. Your computed ρ of 0.89 earlier would show a correlation between the hours of operation and the number of failures if you set your alpha at 0.05, considering that there were six pairs of data. The computed ρ of 0.89 is greater than 0.829, the critical value. This correlation test prepares you to proceed with your assumption if you want to find out why there is a relationship between your two sets of data.

As mentioned earlier, rho is often used to estimate r for the product-moment method. Use Spearman's method only with ordinal data. When you have data *from* interval or ratio measurement scales, you *simply convert* to the ordinal scale by ranking the data.

Self-Test Questions

After you complete these questions, you may check your answers at the end of the unit.

429. Correlation concepts

1. Define correlation.

2. Define the term coefficient of correlation.

3. Which of the following correlation coefficients represents the highest relation: 0, 0.5, –0.87, –0.97, or 0.97?

4. What is a scatter diagram?

5. What does each point on a scatter diagram represent?

6. Refer to the graphic as you match the diagrams in Column B with the most appropriate interpretation listed in Column A. Each answer may be used only once.

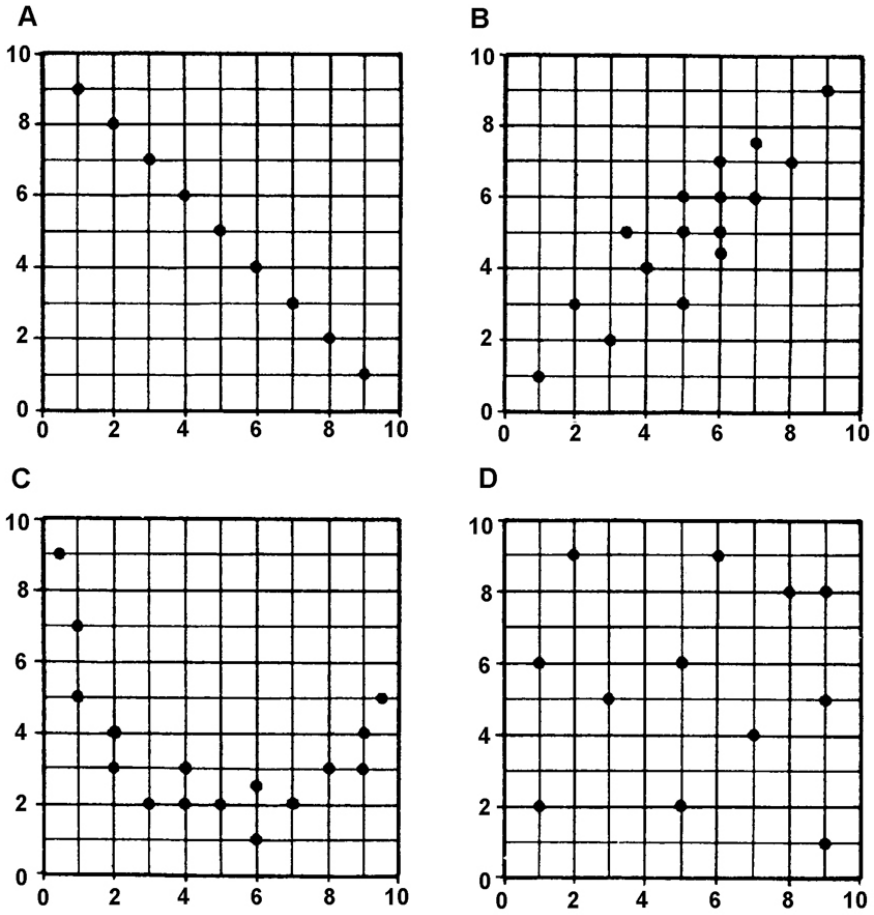


Figure T–1. Scatter diagram (self-test).

Column A	*Column B*
____ (1) Little or no relationship.	a. Scatter diagram A.
____ (2) Perfect negative relationship.	b. Scatter diagram B.
____ (3) High positive relationship.	c. Scatter diagram C.
____ (4) Nonlinear relationships	d. Scatter diagram D.

430. Performing Pearson's product-moment correlation

1. What symbol denotes Pearson's coefficient of correlation?

2. What does Pearson's coefficient of correlation measure?

3. What criteria must a sample meet before Pearson's product-moment correlation techniques can be applied?

4. What does the *minimum* value of r associated with the table of critical values for Pearson's r represent?

5. Compute Pearson's r for the values: $\Sigma X = 9981$; $\Sigma Y = 3687$; $\Sigma X^2 = 12550713$; $\Sigma Y^2 = 1755393$; $\Sigma XY = 4545003$; and $N = 10$.

431. Performing Spearman's rank correlation coefficient

1. What happens to the *extreme* values when Spearman's method of correlation is applied? Why?

2. When should you apply Spearman's method of correlation?

3. What is Spearman's rank correlation coefficient called?

4. What is the relationship between Spearman's correlation coefficient and the coefficient using the product-moment method of correlation?

5. Compute Spearman's rank correlation coefficient using $\Sigma D^2 = 22.2$, and sample of 24.

4–2. Trend Analysis

Several general terms are used with trend analysis. These terms describe the type of trend, type of variations, and method of forecasting.

432. Performing time series analysis

Since forecasts are based on past and present data that is usually represented by a set of observations made at consecutive time periods, these sets of observations are called *time series*. A time series represents the data upon which a trend analysis is performed. It is a systematic arrangement of observed occurrences of data by some constant of time. For example, the following is a time series:

Month	Failures
Jan	3
Feb	5
Mar	7
Apr	9
May	4
Jun	2

The values represented by a time series almost always display some variation. This variation is called a *trend*. Variations of the data in a time series usually fit into one of three categories: secular variations, seasonal variations, or cyclical variations occur in most time series.

Secular trend variation

The term *secular trend* describes the most general and *most* common type of variation in a time series. It describes the general disposition of the data over a long period of time. Usually, you need at least two years of data (on a monthly basis) to determine if a time series has any specific variation. The general forces of nature cause secular trend variation. Depending on the net effect of these forces, a time series may reveal an upward trend, downward trend, or no trend. These are the three ways to describe secular trend variation.

Seasonal trend variation

When observations in a time series are made at intervals shorter than a year (weekly, monthly, or quarterly), they may exhibit seasonal variations. Again, it takes at least two years of data to determine if a variation is seasonal. At bases where there are four definite seasons—fall, winter, spring, and summer—seasonal variation inevitably occurs in much of the maintenance data.

Your maintenance organization may experience many seasonal variations. One, it takes technicians longer to change an aircraft component outside when they are hot or cold than it does when they are comfortable. Two, some types of components fail more often in hot weather than in cold weather. Third, there are usually fewer man-hours available for maintenance in December, June, July, and August than in other months.

Seasonal variations may occur because of reasons other than weather and holidays. For example, local purchase expenditures seem to increase toward the end of the fiscal year, as do the number of temporary duties (TDY). Whatever the case may be, seasonal variations seem to depend on the effect of seasonal influences.

Cyclical trend variation

Variations caused by the nature of the data itself sometimes fall into cycles. Cyclical variation is the *least significant* type of time series variation and has little proof of its existence. Many statisticians believe, however, that all data, over a period of a long time, go through periods of high and low points. Thus, the period of time varies *according to the nature of the data and not on the outside*

forces acting on the data. For instance, an aircraft may experience high systems failures during the first year following production. These types of cycles are normally attributed to the simple passage of time or debugging phase.

Now, knowing that a time series is a set of observations taken over consecutive time periods and that this time series is subject to variation, this is just the beginning of trend analysis. To perform the entire analysis, examine the time series to determine the nature of its trend and then extrapolate to predict future occurrences.

Analysis of linear trends

To perform a trend analysis, you must amass data in the form of a time series, plot the data on graph paper, and then determine whether the data varies secularly, seasonally, or cyclically. If the data vary secularly, employ techniques to determine the direction and magnitude of the trend.

In the straight-line time series, the data, when plotted on regular graph paper, can be approximated fairly well by a straight line drawn through it (fig. 4–11). The line in figure 4–11 was drawn freehand; therefore, it may not be the best line for the data.

To determine the line that best "fits" the data, use one of two mathematical techniques. The first method is the semi-average method; it is much simpler to compute but less accurate. The second method is the least-squares method, which is not only more precise but also more difficult to determine. Always use the least-squares method when extrapolation is to follow the process. Extrapolation may also be used to predict future occurrences when the trend is "linear," or in other words, follows an approximate straight line.

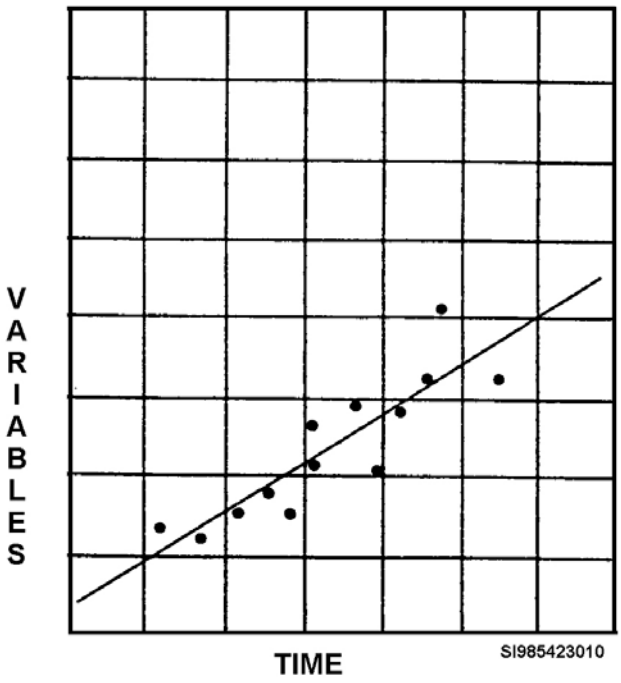


Figure 4–11. Linear trend line.

Rules for linear trend techniques

Most of the restrictions involved with plotting a linear trend line have been alluded to earlier. However, these restrictions have not been stated as a group. There are three general rules to follow when considering a linear trend technique. These rules place certain restrictions on the use of these techniques.

Linear Trend Technique Rules	
Rule	**Consideration**
1	It *must* be *possible* to categorize the data by some time period; preferably by month for maintenance data. This allows the formulation of a time series.
2	The time series that represents the data must be at *least* 24 months in duration. Anything less than 24 months of data is difficult to visualize.
3	The data, when plotted, must have a reasonable *resemblance* to a straight line. This requires judgment on the part of the analyst. Experience in plotting data is needed to know which data display straight line tendencies and which do not.

Semi-average trend line

In the discussion of measures of central tendency, you will learn that the arithmetic mean is a typical value and is representative of a series. If a time series is broken in half, the two equal parts are represented by their respective means. Therefore, a straight line passing through the two means may be considered a rough description of the total trends. This type of trend line is called a semi-average trend line (fig. 4–12 and fig. 4–13). The data used in figure 4–12 is the same as that used for the least-squares method so that you may compare the trend lines determined by the two methods.

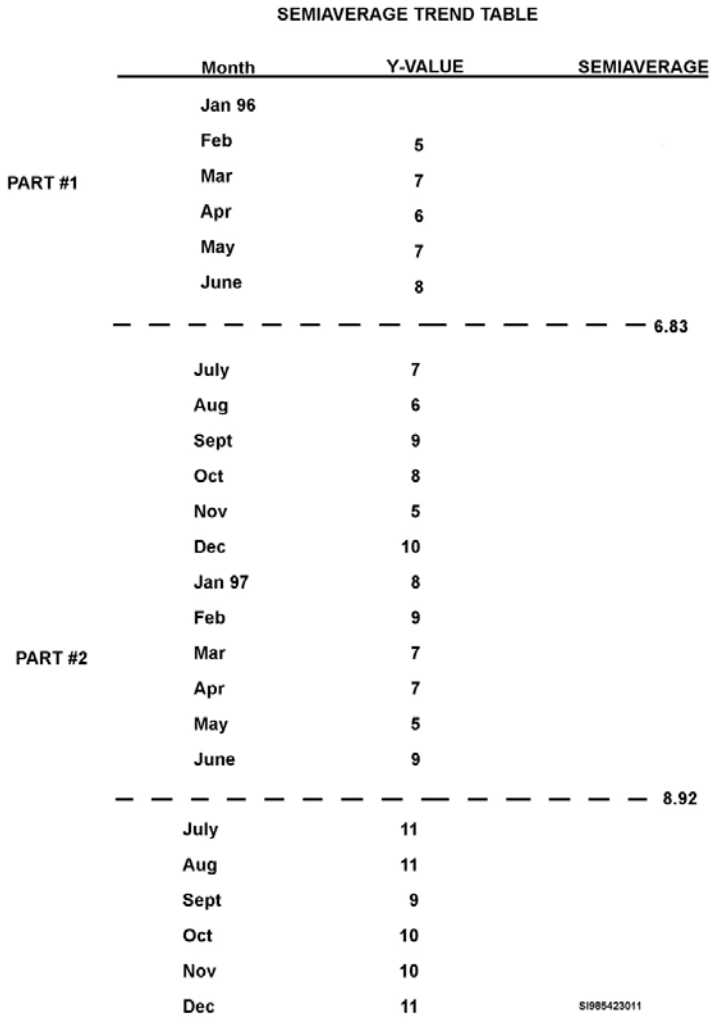


Figure 4–12. Semiaverage trend.

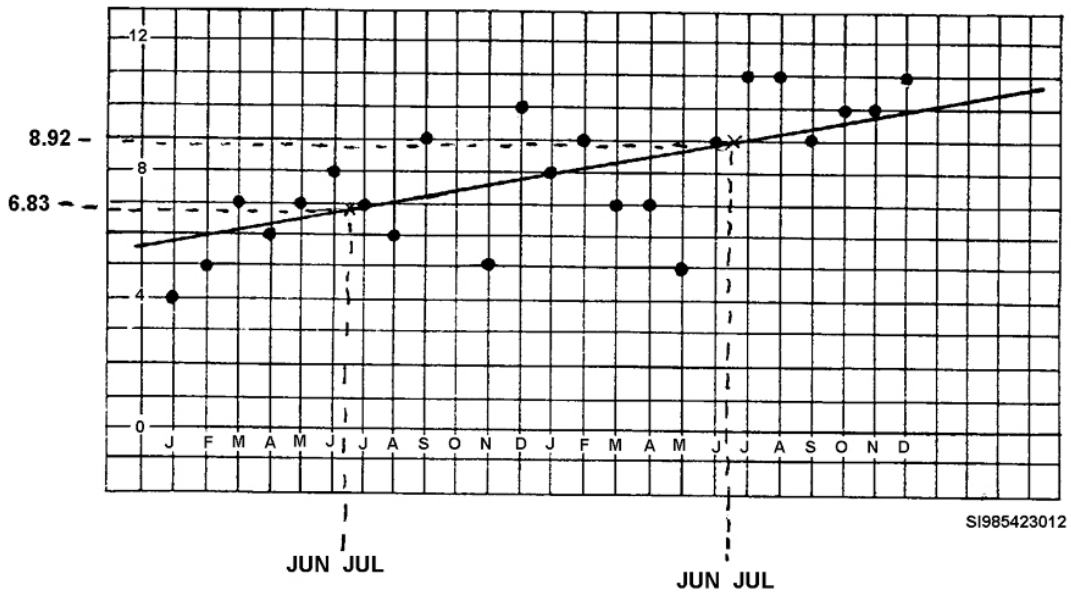


Figure 4–13. Plotted trend data from figure 4–12.

Note in the semiaverage method, the two plots (6.83 and 8.92) are made in the center of each half of the broken time series. The beginning and ending months are *not* used to make the plots because they are *not* considered the average month in the series. When dividing a time series that has an odd number of months, the middle month may be added to each half or omitted from the computations altogether.

Remember, the semiaverage method uses the mean, which is greatly affected by extreme values in a series. When a high degree of accuracy is needed or the data series contain extreme values, then this method is not appropriate. It is not subjective, however, and provides a quick, easy method for many applications.

433. Using the least-squares method

Whenever data *tends* to form a straight line, the least-squares method is the *most popular method* for computing the secular trend of a time series. The computed trend values form a straight centerline through the time series with some of the actual values falling above the least-squares line and some below. The name *least-squares* is derived from the fact that when the amounts of the actual values deviate above or below the trend line, the deviations are measured and squared. The sum of these squared deviations is at a minimum for the line that is called the line of best fit. However, when extremely large deviations are present, the least-squares method may tend to give too much weight to the extremes. Use the T test for outliers to determine whether such extremes should be eliminated from computations.

Odd number of months

There are several methods you can use to compute the least-squares trend line. However, the method presented here is considered the best because it does not require solving complicated equations. Determine the trend values by solving the following basic equation:

$$Y_C = a + bX$$

Where:

Y_C = computed trend value

a = average monthly value

b = slope of the trend line X = variable assigned to a particular month

The formula for “a” is:

$$a = \frac{\Sigma Y}{N}$$

Where:

Y = actual monthly values

N = number of months used

The formula for “b” is:

$$b = \frac{\Sigma XY}{\Sigma X^2}$$

(b may be either positive or negative)

Figure 4–14 illustrates the statistics used to determine the various totals required for an odd number of months.

	X	X^2	Y	XY	Y_C
Jan	–5	25	3.8	–19.0	3.79
Feb	–4	16	4.0	–16.0	
Mar	–3	9	3.8	–11.4	
Apr	–2	4	4.6	–9.2	
May	–1	1	4.5	–4.5	
Jun	0	0	4.1	0	4.546
Jul	1	1	4.8	4.8	
Aug	2	4	4.9	9.8	
Sep	3	9	5.2	15.6	
Oct	4	16	5.0	20.0	
Nov	5	25	5.3	26.5	5.3
	N = 11	110	50.0	16.6	

SI105358144

Figure 4–14. Least-squares trend (odd number of months).

NOTE: For simplicity, only 11 months of data were used.

$$a = \frac{\Sigma Y}{N}$$

$$a = \frac{50}{11} = 4.5455$$

$$b = \frac{\Sigma XY}{\Sigma X^2}$$

$$b = \frac{16.6}{110} = 0.1509$$

$Y_C = a + bX$

$Y_C = 4.545 + (0.1509)(X)$

When using an odd number of months, the midpoint of the time series falls exactly on the central month (assigned an X value of zero). Number the remaining months consecutively from the midpoint by using negative numbers for prior months and positive numbers for later months.

Now, substituting values of X in the equation, you can calculate values of Y_C. Solving Y_C for January, you use the value –5 for X, thus:

$Y_C = 4.5455 + (0.1509)(–5)$

$Y_C = 4.5455 + (–0.7545)$

$Y_C = 3.791$

Solving Y_C for June, using 0 for X, you find $Y_C = 4.5455$.

$Y_C = 4.5455 + (0.1509)\ (0)$

$Y_C = 4.5455 + 0$

$Y_C = 4.5455$

Similarly for November, X = 5 and

$Y_C = 4.5455 + (0.1509)\ (5)$

$Y_C = 4.5455 + 0.7545$

$Y_C = 5$

Even number of months

When using an even number of months, the midpoint falls between two months, as shown in figure 4–15. (**NOTE**: For simplicity, only 12 months of data were used.) Assign the two central months –1 and +1, and number the remaining months in increments of two. Thus, all X values are odd numbers when N is an even number. The sum of the X column is always zero for both an odd and an even number of months.

	X	X^2	Y	XY	Y_C
Jan	–11	121	6.8	–74.8	6.138
Feb	–9	81	4.0	–36.0	
Mar	–7	49	5.5	–38.5	
Apr	–5	25	5.9	–29.5	
May	–3	9	5.0	–15.0	
Jun	–1	1	4.3	–4.3	
Jul	1	1	3.2	3.2	
Aug	3	9	3.8	11.4	
Sep	5	25	2.7	13.5	
Oct	7	49	3.4	23.8	
Nov	9	81	2.2	19.8	
Dec	11	121	1.5	16.5	1.912
	N = 12	572	48.3	–109.9	

SI105358145

Figure 4–15. Least-squares trend (even number of months).

Determine ΣX^2 by squaring each X value and summing the column. Determine ΣXY by multiplying each X value by the corresponding Y value and summing the results. Be careful not to lose track of signs during multiplication and summing. A positive ΣXY (positive slope) indicates an increasing trend; while a negative ΣXY (negative slope) indicates a decreasing trend.

Although Y_C values may be computed for any given month or all months, you only need two points to plot a trend line since the two points determine the line. Normally, for ease of plotting, you'll want to use the two extreme months. It may be wise to compute a third point as a self-check. If the three points do not form a straight line, you made an error in your computation or plotting. If using X values with the same magnitude, but opposite signs to compute bX, add and subtract the product (bX) from "a" since only the sign of bX changes; this makes the computation easier.

Now, substituting values of X in the equation, you can calculate values of Y_C. Solving for Y_C for January, use the value –11 for X, thus:

$Y_C = 4.025 + (–0.1921)(–11)$

$Y_C = 4.025 + 2.1131$

$Y_C = 6.1381$

Similarly for December, X = 11 and

$Y_C = 4.025 + (–0.1921)\ (11)$

$Y_C = 4.025 + (–2.1131)$

$Y_C = 1.9119$

Figures 4–16 and 4–17 are graphs illustrating the data from figures 4–14 and 4–15, respectively, to show the plotted Y_C values and the corresponding trend lines.

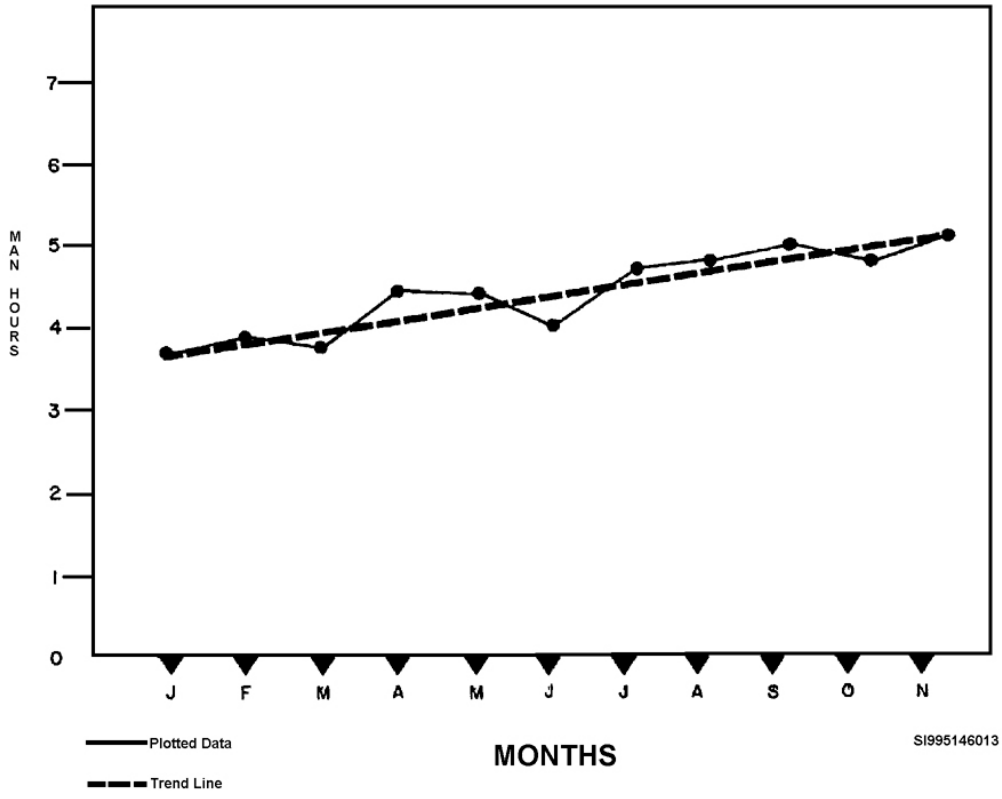


Figure 4–16. Least-squared trend line for odd number of months.

Looking at the trends shown in figures 4–16 and 4–17, there are several things that you should observe. First, the X axis of the charts displays time increments. Second, the Y axis of the charts displays the numerical data. After you compute and plot the Y_C values for the extreme months, you draw a straight trend line through the plotted data to connect the two trend points (computed Y_C values).

Visually, you can readily tell if the trend line is correctly plotted and drawn. There should be an equal amount of plotted data area on each side of the trend line. If there isn't, you made an error and you must recheck your work.

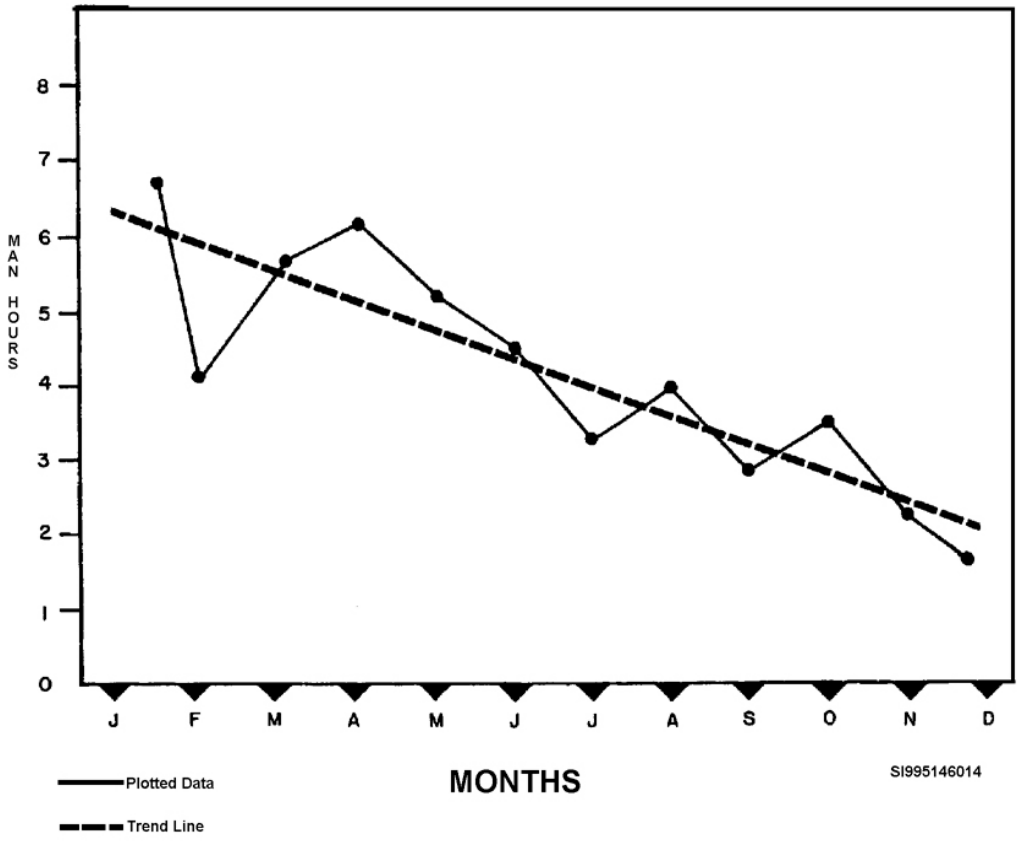


Figure 4–17. Least-squared trend line for even number of months.

434. Nonlinear trends

Because of changes within maintenance procedures, utilization requirements, personnel, and so forth, the direction of a trend changes from time to time. In such cases, a linear trend line will not fit the data; therefore, a method for plotting a curved line is needed. Use a nonlinear trend if your data does not meet linear trend prerequisites. Depending upon the particular time series data, the nonlinear trend line depicts either a gradual leveling off or an actual change in direction in trend. This lesson teaches the use of mathematical formulas to develop nonlinear trends.

Two methods are available for plotting nonlinear trends. The first method, called the *moving average method,* is much simpler to compute yet still provides satisfactory results in most cases. The other, *second-degree parabola method,* is basically the same as the least-squares linear method, with the exception that the former requires more extensive computations. The second-degree parabola is the more precise of the two nonlinear methods offered.

Moving average trend line

The moving average method is a good method for fitting a trend line, and it may also be used for smoothing out seasonal variations, cyclical variations, and so forth. Smoothing out fluctuations with moving averages removes their influence. Thus, the moving average can be used for any part of time series analysis. In general, this method consists of taking the arithmetic mean of a given number of values within a time span and making it the center of the values used. Then the procedure is repeated by dropping the first value used and adding on the value that follows the last figure that was previously used. Thus, the average is "moved" through the time span until the series is exhausted. To obtain the moving average, divide each three-month total by three. Any number of months or time periods may be used to develop the subtotals; for example, three months, four months, five months, and so forth, depending upon the grand total in the series and the fluctuation you desire to eliminate. At any rate, the moving total and moving average *must* be plotted at the center of the time span from which they were drawn.

For example, in figure 4–18 a table of trend data Y was developed for the months of January through November. To get the first three-month averages, you add the first three months' Y values (Jan-Feb-Mar) which are 18+20+25 = 63 and divide it by 3, which is 21. Next, you add the Feb-Mar-Apr Y values (total: 83) and divide again by 3 to get 27.7. You continue with the process until you get to the last three months (Sept-Oct-Nov). The last set of months will give a sum of 169 and a 3-month average of 56.3. Refer to figure 4–19 for a better understanding of how the moving average trend line is developed. Figure 4–19 gives an example of a moving average trend line containing data plotted from figure 4–18.

MOVING AVERAGES

Month	Y	3 Mo Total	3 Mo Moving Average
Jan	18		
Feb	20	63	21.0
Mar	25	83	27.7
Apr	38	115	38.3
May	52	154	51.3
Jun	64	186	62.0
Jul	70	202	67.3
Aug	68	204	68.0
Sept	66	189	63.0
Oct	55	169	56.3
Nov	48		

SI985423018

Figure 4–18. Trend data.

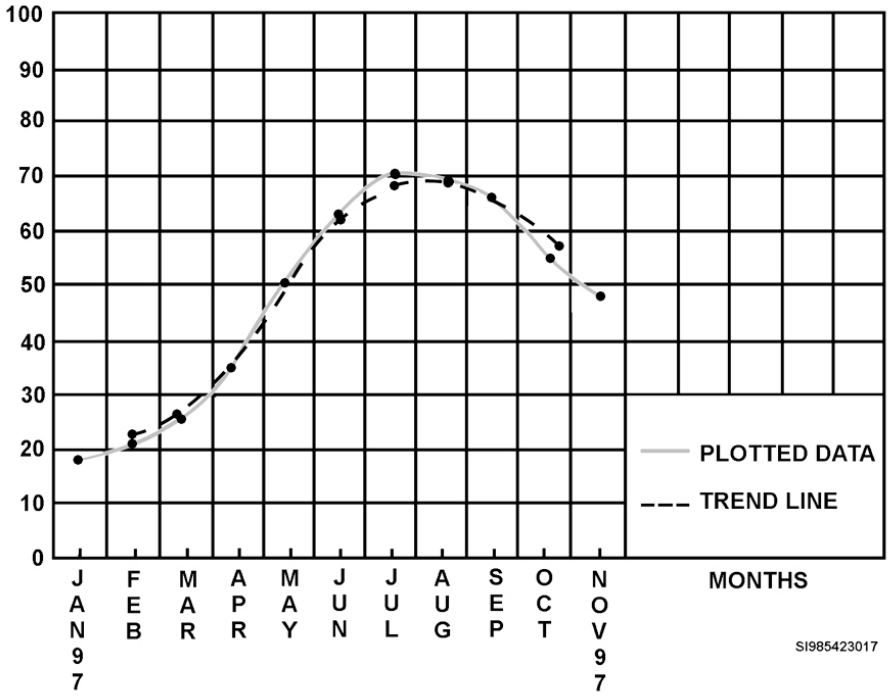


Figure 4–19. Moving average trend.

There are three *disadvantages* of the moving average method:

1. Data at each end of the overall series is *lost*. (If a comparatively short time series is used or if the moving total spans are too large, then losses may be so great as to make the moving average inadvisable.)
2. There is *no* way to determine the exact position of the trend line value for the *most* recent time period.
3. Extrapolation is *not* possible when using this method.

Second-degree parabola trend line

Use the second-degree parabola method to construct a nonlinear trend line. Determine the trend values by solving the following basic equation:

$$Y_C = a + bX + cX^2$$

Where:

Y_C = computed trend value

a = midpoint or average monthly value

b = slope of the trend line

c = determines the degree of change in the slope of trend line

X = variable assigned to a particular month

$$b = \frac{\sum XY}{\sum X^2}$$

Calculate the constant, b, in the same manner as for the least squares linear trend:

$$c = \frac{\left(N\sum X^2Y\right)-\left(\sum X^2\sum Y\right)}{\left(N\sum X^4\right)-\left(\sum X^2\right)^2}$$

$$a = \frac{\sum Y - \left(C\sum X^2\right)}{N}$$

Determine the constant, c, before computing constant a, because c becomes part of the formula for constant a. The constant, c, may be very small, often less than one.

The rules for assigning the X values for an odd or even number of months and calculations of ΣX^2 and ΣXY are similar with the least-squares linear method (shown in figs. 4–14 and 4–15). Computation of the constant "c" necessitates two additional columns for ΣX^4 and ΣX^2Y. Compute the quantity X^4 by raising each X value to the fourth power and summing the results. Calculate the quantity ΣX^2Y by multiplying each X^2 value by the corresponding Y value, and summing the results.

It should not be necessary to compute the trend value for every month. However, enough points must be computed to provide adequate guidance for drawing the trend line. Generally, every third or fourth month should suffice.

Figure 4–20 is an example of the computations for determining nonlinear trend line values. For simplicity, only 11 months of data were used. The same trend data from figure 4–18 is used so that a comparison may be drawn between the "moving average" and the "second-degree parabola" method. Normally 24 months is recommended.

SECOND-DEGREE PARABOLA METHOD USING ODD NUMBER OF MONTHS

	X	X^2	X^4	Y	XY	X^2Y	Y_C
Jan	-5	25	625	18	-90	450	7.391
Feb	-4	16	256	20	-80	320	
Mar	-3	9	81	25	-75	225	35.439
Apr	-2	4	16	38	-76	152	
May	-1	1	1	52	-52	52	53.927
Jun	0	0	0	64	0	0	
Jul	1	1	1	70	70	70	62.855
Aug	2	4	16	68	136	272	
Sep	3	9	81	66	198	594	62.223
Oct	4	16	256	55	220	880	
Nov	5	25	625	48	240	1200	52.031
N = 11		110	1,958	524	491	4,215	

SI985423019

Figure 4–20. Second degree parabola using odd number of months.

As with the least-squares method for linear trends, when you use an *odd* number of months, the midpoint of the time series falls exactly on the central month, and this month is assigned an X value of zero. Number the remaining months consecutively from the midpoint by using negative numbers for prior months and positive numbers for later months.

When you use an *even* number of months, the midpoint falls between the two central months. For the two central months, you assign X values negative one and positive one, and you use a difference of two when numbering remaining months. Thus, all X values are odd numbers when N is an even number. The sum of the X column will always be zero for both an odd and an even number of months.

Using the values from figure 4–18, calculate a, b, and c as follows:

$$b = \frac{\sum XY}{\sum X^2} \qquad b = \frac{491}{110} \qquad b = 4.464$$

$$c = \frac{\left(N\sum X^2Y\right) - \sum X^2 \sum Y}{\left(N\sum X^4\right) - \left(\sum X^2\right)^2}$$

$$c = \frac{11(4215) - 110(524)}{11(1958) - (110)^2} \qquad c = -1.195$$

$$a = \frac{\sum Y - C\left(\sum X^2\right)}{N} \qquad a = \frac{524 - (-1.195)110}{11} = \frac{524 + 131.45}{11} = \frac{655.45}{11} = 59.586$$

$$Y_c = a + bX + cX^2$$

$$Y_c = 59.586 + 4.464X + (-1.195)X^2$$

Substituting the values of X and X^2 from figure 4–20 for January and doing the arithmetic, $Y_C = 7.391$. Since the trend line is a curve instead of a straight line, the equation for Y must be solved for each trend point to be plotted.

Figure 4–21 illustrates a second-degree parabola trend line using the data from figure 4–20.

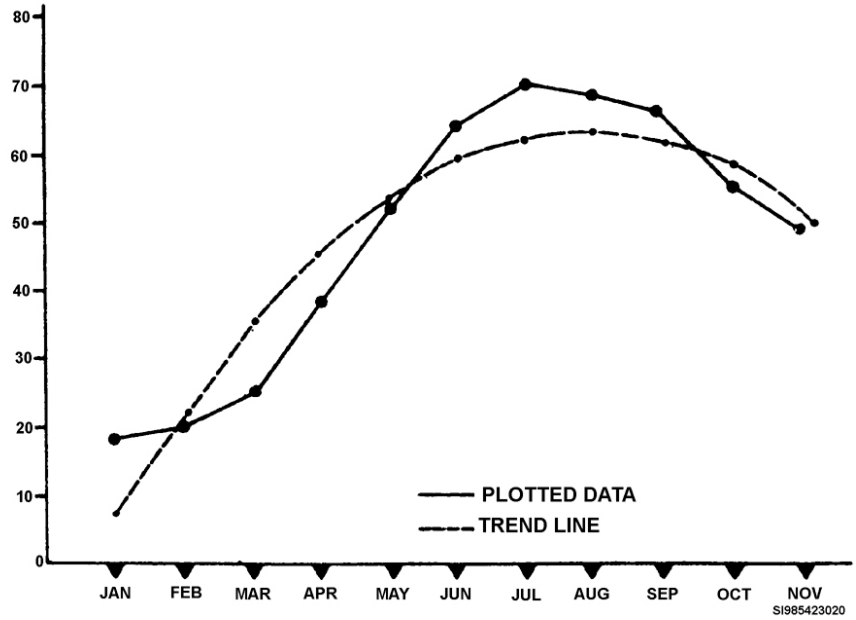


Figure 4–21. Second-degree parabola trend line.

The nonlinear trend line is plotted by using the second-degree parabola method of computation. As you can see, this method gives a more realistic picture of the trend the data is making than a linear trend line would if plotted on this same chart. There is really no way a straight line can be placed to fit this data.

435. Seasonal trends

Use the percent-of-yearly-total method to establish an *index* of *seasonal* variations. This method is a simple means of measuring seasonal variations. As the name implies, the resulting index shows the average percent of the year's total that can be expected to occur during each particular month. In other words, the overall total for the entire year equals 100 percent. If there are no variations, each month will account for 8 $^{1}/_{3}$ percent (100/12) of the year's total. However, when variations exist, there may be certain months or periods when the respective percentages are consistently higher or lower than others. As an example, overtime may increase during certain periods because of leave or increased commitments. The percent-of-yearly-total method shows which months are above or below the overall average.

To prevent distortion by one-time random occurrences, use at least two years of data. This is assuming no major changes in procedures, and so forth, have occurred during the time span. Normally, data is compiled for the calendar year starting in January; however, any two 12-month period can be used.

You do not need a formula to use the percent-of-yearly-total method. Simply total all data for each month individually (for the two years), and add the 12 monthly totals to obtain a grand total for the entire two-year period. Then, find each month's total percentage of the grand total. Compute the percentages by dividing each of the monthly totals by the grand total. The 12 percentages should total approximately 100 percent. Figure 4–22 illustrates the computation of seasonal indexes, using the percent of yearly total method.

Monthly	Last Year	This Year	Monthly Total	Percent
Jan	9.9	11.9	21.8	8.97
Feb	10.2	12.1	22.3	9.18
Mar	10.4	11.8	22.2	9.14
Apr	9.8	10.5	20.3	8.35
May	8.9	10.3	19.2	7.90
Jun	8.5	10.1	18.6	7.65
Jul	8.0	10.0	18.0	7.41
Aug	8.1	10.0	18.1	7.45
Sep	8.5	10.1	18.6	7.65
Oct	9.1	11.9	21.0	8.64
Nov	9.5	11.8	21.3	8.77
Dec	9.6	12.0	21.6	8.89
Totals	**110.5**	**132.5**	**243.0**	**100.00**

Figure 4–22. Seasonal trend.

For visual analysis, graph your indexes, as shown in figure 4–23, when analyzing seasonal data. The 8 1/3 percent centerline serves as a reference point to indicate which months are above or below the overall average.

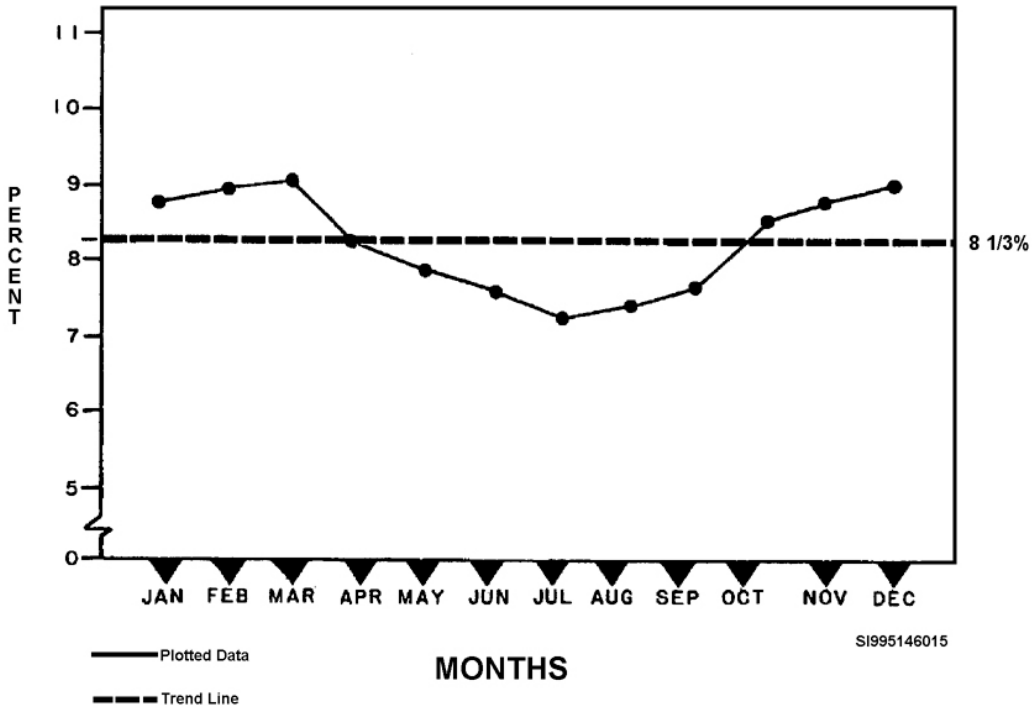


Figure 4–23. Seasonal index.

The example presented should have given you a basic understanding of the percent-of-yearly-total method used for establishing an index of seasonal variations.

Self-Test Questions

After you complete these questions, you may check your answers at the end of the unit.

432. Performing time series analysis

1. Define a time series.

2. Define the three types of variation that normally occur in a time series.

3. What are the three rules you should follow when considering a linear trend technique?

4. What measure of central tendency does the semi-average use?

433. Using the least-squares method

1. Write the three formulas *required* to compute a least-squares trend line.

2. Using figure 4–15, compute Y_C for May and September.

3. When calculating a 24-month trend from Jan 2002 to Dec 2003, what X value is assigned to Nov 2002, Dec 2002, Jan 2003, and Feb 2003?

4. When calculating a 23-month trend from Feb 2002 to Dec 2003, what X value is assigned to Nov 2002, Dec 2002, Jan 2003, and Feb 2003?

434. Nonlinear trends

1. When would you use a nonlinear trend?

2. Which of the two nonlinear trend methods is the *most* precise?

3. What are the three *disadvantages* of the moving average method?

4. What is the *basic* equation for the second-degree parabola trend line?

435. Seasonal trends

1. When totaling the monthly percentages, the yearly total equals what value?

2. Each month accounts for what percent of the year’s total?

4–3. Extrapolation

To predict from a trend, you must extend the trend from the past to the future through extrapolation. In this section, we discuss the methods of extrapolation that apply to linear and seasonal trends. We also talk about Kendall's test to check the significance of trends.

436. Performing extrapolation of linear and seasonal trends

In regression analysis, extrapolation is the process of constructing new data points beyond the set of known data points. For linear and seasonal trend lines, this means creating a tangent line at the end of the know data and extending it beyond that limit. In this lesson, we will discuss extrapolation of linear and seasonal trends by mechanical and mathematical methods.

Extrapolation of linear trends

The use of statistical methods in the prediction of secular trends consists *primarily* of the measurement of events in the *past*. A very basic method of predicting a trend is by *extrapolation* (e.g., the extension of a fitted trend line into the future for the number of months for which a prediction is desired).

The two methods of extrapolation are *mechanical* and *mathematical*. The mechanical method is an extension of the existing trend line. The mathematical method requires computation to determine the rate of change, which is either subtracted or added to the most recent computed trend value (Y_C).

Mechanical method

The mechanical method of using a trend for prediction purposes is commonly referred to as extrapolation. This method refers to the expression of the linear trend line past its present limits. Extrapolation is an easy method of predicting linear data, but it *cannot be used with nonlinear trends*. Let's say a least-squares trend line was developed with data from the table shown in figure 4–24. We will develop a trend line using the least-squares method.

We use the formula for the least-squares method to find at least two points (usually near the beginning and end of the time period) that can be plotted and joined by a straight line. Recalling the formula:

$$Y_C = a + bX$$

Where:

Y_C = computed trend value

a = constant

b = the slope of the trend line

X = the assigned variable

$$a = \frac{\sum Y}{N}$$

Where:

$$b = \frac{\sum XY}{\sum X^2}$$

ΣY = sum of actual values and N is the number of values in the time series.

The constant, a, is always positive. The constant, b, is computed:

Where:

X = the value assigned to the data along the time scale.

Remember that the constant, b, may be either positive or negative.

Let's use the data in figure 4–24 to calculate Y_C, the trend, using the least-squares method. If the data were plotted, the points would appear to fall in a straight line and seem to have a secular trend variation.

LEAST SQUARES TREND TABLE

Month	X	X^2	Y	XY	Y_c
Jan 96	-23	529	4	-92	5.483
Feb	-21	441	5	-105	
Mar	-19	361	7	-133	
Apr	-17	289	6	-102	
May	-15	225	7	-105	
Jun	-13	169	8	-104	
Jul	-11	121	7	-77	
Aug	-9	81	6	-54	
Sep	-7	49	9	-63	
Oct	-5	25	8	-40	
Nov	-3	9	5	-15	
Dec	-1	1	10	-10	
Jan 97	1	1	8	8	
Feb	3	9	9	27	
Mar	5	25	7	35	
Apr	7	49	7	49	
May	9	81	5	45	
Jun	11	121	9	99	
Jul	13	169	11	143	
Aug	15	225	11	165	
Sep	17	289	9	153	
Oct	19	361	10	190	
Nov	21	441	10	210	
Dec	23	529	11	253	10.267
TOTALS		4600	189	477	

SI985423013

Figure 4–24. Least-squares trend table.

Since the number of the months in the time series is *even*, then the X values are assigned like those in figure 4–24.

The two middle months are assigned values of –1 and 1, respectively. Then, the values increase in magnitude by 2s; the months prior to the –1 being negative and those after the 1 being positive.

Figure 4–24 consists of several columns for which you must make computations in order to attain Y_C. The X^2 and XY columns on figure 4–24 are self-explanatory. In order to find Y_C, the totals of columns X^2Y, and XY are needed. In our example, ΣX^2 is 4,600, ΣY is 189, and ΣXY is 477. Pay particular attention to the sign of the ΣXY, as this has a great deal to do with the direction of the trend.

To reemphasize, to find Y_C, solve the formula:

- $Y_C = a + bX$, which means you need to find a and b. First find a, which is the simple arithmetic mean of the Y values:
- Now substitute a and b in the formula for Y_C at any particular month (represented by the X values). The formula for Y_C becomes: $Y_C = 7.875 + .104(X)$ for this particular situation.

As stated earlier, two points are necessary to draw a straight line through the data. It is usually best to find these points at the two extreme ends of the data. Therefore, determine Y_C for Jan 96 and Dec 97. The two X values, then, are –23 and 23.

Solving for Y_C for the two values:

$$Y_C \text{ (Jan 96)} = 7.875 + .104(-23)$$
$$= 7.875 + (-2.392)$$
$$= 5.483$$

and

$$Y_C \text{ (Dec 97)} = 7.875 + .104(23)$$
$$= 7.875 + 2.392$$
$$= 10.267$$

Figure 4–25 illustrates the plotted data and the completed trend line for the information in figure 4–24.

To plot the trend line in figure 4–25, you had to solve the equation:

$$Y_C = a + bX$$

This required the formulation of a trend table (fig. 4–24), determining the constants a and b, and solving for Y with two values of X. To be useful, the trend line should be drawn on graph paper.

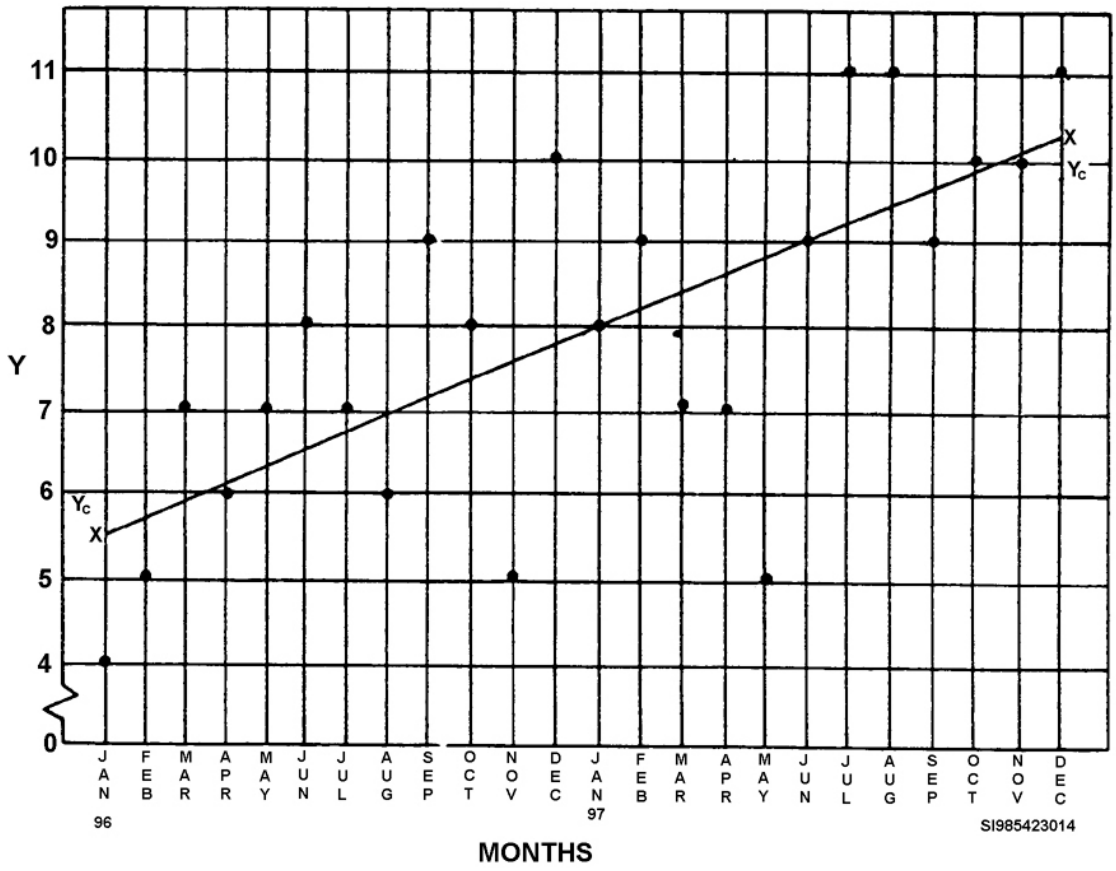


Figure 4–25. Plotted trend data from figure 4–24.

Figure 4–26 displays the extrapolation of the trend line (solid line) from figure 4–25. The data used to derive this trend line started in January 1996 and ended in December 1997.

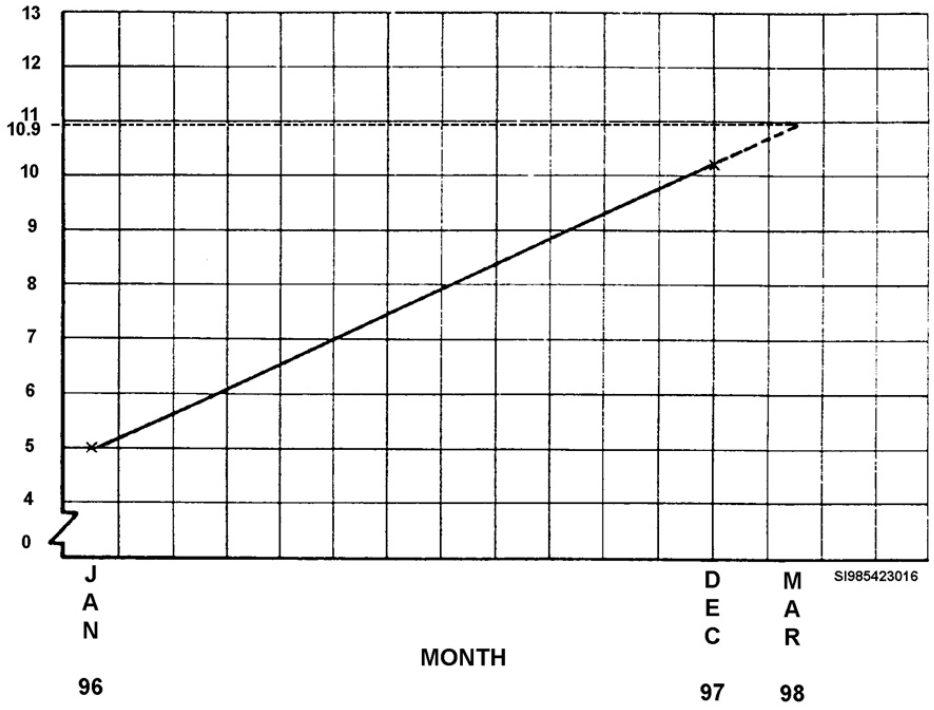


Figure 4–26. Extrapolation of trend line.

If you wish to predict the Y value that should occur for the month of March 1998, you simply have to extrapolate this trend line for three months. The extrapolation is indicated by the dotted continuation of the solid trend line (using a straight ruler and pen) obtained from figure 4–25. To determine the Y value for March 1998, read horizontally across to the Y-axis of the graph and extract the numerical value from the axis. The trend in figure 4–26 indicates that for March the value is approximately 10.9. Like most statistical methods, however, mechanical extrapolation has some limitations.

Mathematical method

The mathematical method uses the least-squares method to determine the trend value Y_C and computes for the rate of change of the trend value. When you look for the rate of change, you're looking for the amount each month would increase or decrease with the rate of change value.

Look again at figure 4–14. Using its Y_C values, three facts can be stated about the trend of data. First, the presence and direction of the trend are indicated. The trend is increasing since it went from 3.79 to 5.3 in an eleven months' span.

Second, the *rate of change* can be calculated. The rate of change (in percent) was computed by using the first Y_C plot of 3.79 as the base or 100 percent. This was divided into the last Y_C plot of 5.3 to yield a value of 1.3984. Subtracting the 100 percent base and changing this to a percent gives a 39.84 percent overall increase for the time span. Dividing this increase (39.84 percent) by 10 months (n–1) gives a 3.98 percent gain each month. We divided by 10 months instead of 11 because data changed only for the last 10 months. The first month was our base and changes were noted in the following months.

Third, the increase can be computed in terms of the units of Y_C rather than in percent. This was computed by multiplying the first Y_C plot of 3.79 (100 percent base) by the percent rate of change per month of 0.0398. This computes out to 0.1508. Since the units of Y_C are man-hours, it can now also be stated that required man-hours or a particular job is increasing at a rate of 0.1508 man-hours per month. The formula for computing the rate of change and extrapolation is presented next.

Rate of change:

$$Y_{C(l)} = \text{base } Y_C \text{ value}$$

$$Y_{C(n)} = \text{last } Y_C \text{ value}$$

Overall increase or decrease during time span:

$$\frac{Y_{C(n)}}{Y_{C(1)}} - 1$$

Rate of increase or decrease per month:

$$\frac{\dfrac{Y_{C(n)}}{Y_{C(1)}} - 1}{n - 1}$$

Units or man-hours of increase or decrease:

$$\frac{\dfrac{Y_{C(n)}}{Y_{C(1)}} - 1}{n - 1} \times Y_{C(1)}$$

It should be pointed out that accurate forecasting is extremely difficult. Often, there is no logical reason for assuming that a trend will continue for any fixed time in the future. You must be very specific in your wording. For example, “If the trend line continues at its present slope, we predict that the trend line will be located at an approximate number of months, if everything remains constant.”

Extrapolation:

$$\frac{\dfrac{Y_{C(n)}}{Y_{C(1)}} - 1}{n - 1} \times Y_{C(1)} \times \text{desired number of months} + Y_{C(n)}$$

Limitations of extrapolation

There are two general limitations to extrapolation. The first applies only to mechanical extrapolation, while the second applies to both mechanical and mathematical extrapolation.

The first limitation is based on mechanical extrapolation being performed with a straight edge and a pencil and is then judged with the human eye. Because of the unknown exactness of the straight edge, the width of the pencil and the failings of the human eye in regard to discriminating judgments, the mechanical method of extrapolation is subject to technical error.

The second limitation applies to any prediction of the future that is based on the past. In predicting the future in this way, one must assume that the forces that acted on the data of the past will continue to act on it in the future at the prevailing rates of the past. Such an assumption cannot be guaranteed. Therefore, it is wise when using extrapolation—or any other method of predicting the future from past events—to make the prediction in words such as the following: “We predict that, if past data is an indicator of future occurrences, the following should be true: In March, the value of Y should be about 10.9.”

Extrapolation of seasonal trends

Let's use figure 4–27 to illustrate seasonal extrapolation. To predict next year's B–52 recovery time, subtract the total of 110.5 for last year from the 132.5 total for this year, for an increase of 22. Then add 22 to the 132.5 for this year since there was an increase. (Likewise, if there had been a decrease, you would have subtracted the difference from the most current year.) Now, a total of 154.5 can be extrapolated or forecast for next year, if the B–52 recovery time continues to increase the *same* as in the past. Multiply the projected total of 154.5 by the monthly percentages starting with January of 8.97 (or a rate of .0897), for a predicted January recovery time of 13.9. The total of the 12 monthly figures should about equal the projected 154.5 total for next year.

Monthly	Last Year	This Year	Monthly Total	Percent	Extrapolation
Jan	9.9	11.9	21.8	8.97	13.9
Feb	10.2	12.1	22.3	9.18	14.2
Mar	10.4	11.8	22.2	9.14	14.1
Apr	9.8	10.5	20.3	8.35	12.9
May	8.9	10.3	19.2	7.9	12.2
Jun	8.5	10.1	18.6	7.65	11.8
Jul	8.0	10.0	18.0	7.41	11.4
Aug	8.1	10.0	18.1	7.45	11.5
Sep	8.5	10.1	18.6	7.65	11.8
Oct	9.1	11.9	21.0	8.64	13.3
Nov	9.5	11.8	21.3	8.77	13.5
Dec	9.6	12.0	21.6	8.89	13.7
	110.5	132.5	243.0	100.00	

SI105358147

Figure 4–27. B–52 recovery time.

437. Performing Kendall's test for significance of a trend

Kendall's test is a statistical method used to test the significance of a linear trend. It uses ranks and develops a statistic called an S statistic. It will not tell what type of trend line best fits the data, but it will tell whether the trend is significantly increasing or decreasing. Data used in Kendall's test must be from a scale that can be ranked. Each set of data should also contain at least 24 values in a series. Refer to figure 4–28 as the steps for Kendall's test are explained.

1. The first step is to set the data in chronological order (fig. 4–28, column A). In other words, set the data in one column from the oldest to the present. For simplicity, we used only 11 months in this example. The same procedures apply when you use two or more years.
2. After setting the data in chronological order, start ranking the data. Give rank 1 to the smallest item, rank 2 to the next smallest, and so on. If there are any ties, average the ranks of these tied values. In our example there are two tied values. They occupy ranks one and two. The average rank (1.5) is assigned to each of these tied values. Rank 3 is assigned to the next smallest value of 3.8. Continue this procedure until all data items are ranked.
3. Start with the rank on the first row and determine how many ranks have a value above the rank of 1.5. Including the ranks above, there are nine. Take the second row (Feb value, 3.8) with rank 3 assigned and decide how many ranks fall above. When counting the ranks with values above a given month's rank value, *do not* consider any monthly rank that has already had its column value annotated. Taking the third row (Mar value, 3.7) with rank 1.5 assigned, determine how many ranks subsequently fall above. Notice that January and February are not used in determining how many ranks fall above the March rank, 1.5. Eight ranks fall above.

A	B	C	D	E	F
Month	Man-hours	Rank Values Numbers	Rank Value Above	Rank Value Below	Above-Below Difference (Column d minus E)
Jan	3.7	1.5	9	0	9
Feb	3.8	3	8	1	7
Mar	3.7	1.5	8	0	8
Apr	4.3	7	4	3	1
May	4.2	6	4	2	2
Jun	3.9	4	5	0	5
Jul	4.4	8	3	1	2
Aug	4.1	5	3	0	3
Sep	4.6	10	1	1	0
Oct	4.5	9	1	0	1
Nov	4.7	11	0	0	0
					S = 38

SI105358148

Figure 4–28. Ranking data.

Follow this same procedure to determine the ranks that subsequently fall above any specified rank. Continue until you have determined all ranks.

4. After listing the ranks that fall above a given value, use the same process to determine the ranks that fall below each value (column E).
5. In this step, compute the difference between the above and the below ranks by subtracting column E from D and then sum these values. The sum of these values (differences) equals Kendall's statistic S (column F). When obtaining the differences, pay close attention to positive and negative signs of S. A negative S indicates a decreasing trend, and a positive S indicates an increasing trend.
6. In this example, Kendall's S is 38 (positive) and indicates an increasing trend. There were no negative differences in this example, but negative data will occur in some instances.
7. Now that you have determined the value S, find the standard deviation of S. Looking at the formula, you can readily see that only one unknown is required. The sample size, n:

$$\sigma_S = \sqrt{\frac{n(n-1)(2n+5)}{18}}$$

Since the standard deviation of S is *dependent on the sample size alone*, figure 4–29 gives the standard deviation of S for sample sizes between 11 and 30.Since you do not have to compute the standard deviation of S, simply extract σ_s from figure 4–29 based on the particular sample size used. In this case, the sample size (n) was 11. So the standard deviation of S is 12.845 (from the table).

8. You learned that you can take a standard deviation and the mean of a series having a normal distribution and determine where a given value falls in the series by using the normal curve area table. In the case of a changing trend, the amount of change, as shown by the S statistic, represents the distance that a value is located from the mean in its distribution. This reasoning is based on the theory that if a distribution is normal, the average increase or decrease in the trend of data is zero, or unchanging.

n	σ_s
11	12.845
12	14.583
13	16.391
14	18.267
15	20.207
16	22.211
17	24.276
18	26.401
19	28.583
20	30.822
21	33.116
22	35.464
23	37.864
24	40.315
25	42.817
26	45.369
27	47.969
28	50.616
29	53.310
30	56.051

Figure 4–29. Standard deviations of S.

Let's compute a Z value (deviation value) for the S statistic to use with the normal curve table. The formula for the Z deviate is:

Where:

$|S|$ = Kendall's "S" which is the absolute value of sum of the above/below rank differences

σ_s = Standard Deviation of Kendall's "S" or $\sqrt{\frac{N(N-1)(2N+5)}{18}}$

$$Z_S = \frac{|S|-1}{\sigma_s}$$

Or in this case:

$$Z_S = \frac{38-1}{12.845}$$

$$Z_S = \frac{37}{12.845}$$

$$Z_S = 2.88$$

The formula uses the absolute value ($|\ |$) of the S statistic. Absolute value means that regardless of whether the K statistic is positive or negative; make it positive. In other words, ignore the sign and use the whole number.

9. To complete the process now that the Z deviate, Z_S, has been computed and based on Kendall's S, use the "area in one tail of the normal curve" sampling distribution. Looking at figure 4–30, read down the Z column and then across until you find the Z value of 2.88. Reading from the body of the table, you'll find that 2.88 represent a probability of 99.6 percent that the trend is significant. You may recall that the normal curve area table represents only half of a normal distribution; therefore, multiply the table value 0.4980 × 2 to obtain 0.996 or 99.6 percent. In turn, you may conclude that the probability is less than 0.4 percent that the change in the trend is due to chance cause.

A sound knowledge of normal distributions is needed when using this trend measurement method. You should now have a basic understanding of Kendall's test for measuring the significance of a linear trend. Use this technique to measure gradual changes in an item that appears to be in control when using other measures. Normally, a probability of significance of at *least* 95 percent or higher is *recommended* before further investigation of the problem is warranted. However, as stated earlier, evaluate each trend on its own merits such as its importance to the mission and the amount of time and people available to conduct the investigation.

Z	0.0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990									
3.2	.4993									
3.3	.4995									
3.4	.4997									
3.5	.4998									
4.0	.4999									

NORMAL CURVE AREA TABLE

Figure 4–30. Normal curve table.

Self-Test Questions

After you complete these questions, you may check your answers at the end of the unit.

436. Performing extrapolation of linear and seasonal trends

Use the following information to answer questions 1 through 3:

The mean time between failures Y_C rates for a 24-month span (Jan 2002 to Dec 2003) rose from 3.75 (Y_1) to 7.9 (Y_n).

1. What was the overall increase?

2. What was the rate of increase per month?

3. What is the mean time between failures rate of increase?

4. If an aircraft system failed 104 times in 2002 and 112 times in 2003, how many failures are expected in 2004 based on the seasonal index?

5. If an aircraft system failed 113 times in 2002 and 110 times in 2003, how many failures are expected in 2004 based on the seasonal index?

437. Performing Kendall's test for significance of a trend

1. Define Kendall's S test.

2. What happens when there are ties within the values being ranked?

3. What is indicated when the computed value of S is positive? Negative?

4. Using n = 15 and S = 24, determine the probability that chance causes are affecting the trend. Use figures 4–29 and 4–30.

5. Using a probability significance of 95 percent, determine whether the trend in question 4 should be investigated.

4–4. Regression Analysis

Correlation analysis looks at the existence of a relationship, while regression analysis measures the validity of the relationship of those data. When you have established that a linear relationship exists between two sets of data through correlation analysis, you're ready to make a strong prediction of the trend by using regression analysis. You use linear regression to make a calculated prediction. Regression analysis uses correlation techniques that take a predetermined relationship and predict future occurrences. For example, if you can predict sortie aborts by the number of engine failures for a fleet of aircraft, you can predict that if the number of engine failures increase, then the number of sortie aborts will increase as well.

In this section we cover the construction of the line of regression and use the formula. Then we show you how to determine the data dispersion about these lines by using the standard error of the estimate. We conclude by predicting data occurrences using the standard error of the estimate.

438. Computing the line of regression

Regression means to go back. Since there is no perfect correlation, we predict using a regression equation or formula where the estimated future occurrence is predicted based on the average values of the past data. A line of regression indicates the trend of the data relationships. The line of regression is calculated from the data in your distribution. Since you are using the actual data occurrences, your line is based on what happened in the past. It provides you with an average value. The line of regression calculation allows you to statistically draw a trend line through a pre-plotted scatter diagram of data.

Steps

Regression analysis has three steps. The first step is to construct a line of regression to show the trend of the data. The second step is to calculate and use standard error of the estimate. This shows the way the data is dispersed or scattered about the line of regression. The third step is to predict by using the line of regression and standard error of the estimate.

Constructing the line of regression

The line of regression is a line that best shows the trend of correlated data. The technique used to graph this line is the least-squares method.

With the points scattered all over the chart on the scatter diagram, the least-squares method makes it easier for you to draw the line in the middle of the area where the points seem to concentrate. When the line is drawn in the middle, all points surrounding the line will be at the shortest distance from the line.

Let's see how to draw the line of regression. The formula for the least-squares line is

$$Y = a + bX$$

Where:

Y = Computed value of Y for a given value of X

a = Point where line crosses the Y axis

b = Slope of the line

X = A selected value

$$b = \frac{N(\sum XY) - (\sum X)(\sum Y)}{N(\sum X^2) - (\sum X)^2}$$

$$a = \frac{\sum Y - b(\sum X)}{N}$$

NOTE: The slope of a line shows rate of increase in Y with increase in X (ratio of distance in vertical direction to distance in horizontal direction).

Before the least-squares line can be plotted, you need to determine the values of "a" and "b."

Let's look at the table of values for Set X and Set Y (fig. 4–30) to see how to determine "a" and "b." In the table we have the values for the necessary sums. For this data, N = 10, ΣX = 205, ΣY = 153.5, ΣX2 = 5725, and ΣXY = 3890.5. Upon substituting these values into the formula for "b" we have

$$b = \frac{10(3890.5) - (205)(153.5)}{(10)(5725) - (205)^2} = \frac{38905 - 31467.5}{57250 - 42025} = \frac{7437.5}{15225} = .4885$$

$$a = \frac{153.5 - .4885(205)}{10} = \frac{153.5 - 100.1425}{10} = \frac{53.3575}{10} = 5.3358$$

By the value of a, you can see that the line crosses the Y-axis at 5.34 (where X = 0) and the slope of the line b is .49.

NOTE: The values are rounded to the nearest hundredth.

Now we can plot the line of regression. It takes two points to draw a straight line. So we pick two X values and compute one Y value for each X value. Usually the X values are selected somewhat apart from each other so a more accurate line of regression can be drawn. Our X-axis goes from 0 to 40 so we will pick one X value of 5 and another X value of 30. By substituting the X value of 5 into the regression line formula together with a = 5.34 and b = .49, we get:

Y = a + bX

Y = 5.4 + (.49) (5)

Y = 5.34 + 2.45

Y = 7.79

So for the selected X value of 5, we get a Y value of 7.79. Similarly using X = 30, a = 5.34 and b = .49, we get:

Y = a + bX

Y = 5.34 + (.49) (30)

Y = 5.34 + 14.70

Y = 20.04

So for the selected X value of 30, we get a Y value of 20.04.

We now have two points: Point 1 where X = 5, Y = 7.79; and point 2 where X = 30, Y = 20.04) through which to draw the line of regression. We plot the points on the graph (represented by two diamond points). Draw a line from one end of the graph to the other that goes through these two points as illustrated in figure 4–31.

So, to plot the line of regression, you determine the values of "a" (the point where the line crosses the Y-axis) and "b" (the slope). These values, together with two arbitrary X values, were substituted in the formula, "Y= a + bX." Two Y values were obtained from this. As a result, we then had two points through which to draw the line of regression.

Plotting the points

Using the same line of regression graph (fig. 4–31), we plot the points from the table in figure 4–30 that represent each pair of X- and Y-values (pairs 1–10). The 10-round points show where each pair of data stands in relation to the average value. The closer the points are to the line of regression, the better the data represents the correlation of the two sets.

Pair No.	X	Y	X^2	Y^2	XY
1	4	7.0	16	49	28
2	6	8.0	36	64	48
3	7	9.5	49	90.25	66.5
4	12	11.0	144	121	132
5	19	13.5	361	182.25	256.5
6	23	16.5	529	272.25	379.5
7	26	20.0	676	400	520
8	32	21.0	1024	441	672
9	37	22.5	1369	506.25	832.5
10	39	24.5	1521	600.25	955.5
	ΣX	ΣY	ΣX^2	ΣY^2	ΣXY
	205	153.5	5725	2726.25	3890.5

SI015423030

Figure 4–31. Values for set X and set Y.

439. Computing the standard error of the estimate

The line of regression is useful in predicting one value when the other value is given. However, when predicting a value based on the line of regression, you are actually using an average. You have more or less averaged a series of points using the least-squares method, and you have drawn a line representing this average. From this "average," you cannot tell how the data was dispersed about the line. A measure of dispersion is necessary to get a realistic picture of the data. In correlation, the measure of dispersion about the line of regression is the standard error of the estimate.

Since there can be two lines of regression (X on Y or Y on X), there also can be two standard errors of the estimates. If you use an X on Y regression line, the standard error of the estimate is S_{XY}. If you use a Y on X regression, the standard error of the estimate is S_{YX}. The S in S_{XY} and S_{YX} shows the measure of dispersion just like the S for standard deviation and $\bar{S}_x$ for standard error of the mean. The subscripts $_{XY}$ and $_{YX}$ indicate which line of regression is referred to.

The standard error of the estimate is based on the normal curve area table shown on the table of FO 1A, just like standard deviation and standard error of the mean. One standard error of the estimate measured off both sides of the line of regression equals about a 68 percent confidence interval. Similarly, two standard errors of the estimate have approximately 95 percent of the data, and three standard errors of the estimate contain approximately 99 percent of the data.

The formula for computing the standard error of the estimate for a Y on X regression line is:

$$S_{YX} = S_Y\sqrt{1-r^2}$$

Where:

S_Y = standard deviation of the Y values

r = coefficient of correlation

S_{YX} = standard error of the estimate

As you probably realize, the standard deviation hasn't been calculated yet. The standard deviation formula for the Y values is:

$$S_Y = \sqrt{\frac{\Sigma Y^2 - \frac{(\Sigma Y)^2}{N}}{N-1}}$$

Where:

Y = actual Y values

N = number of Y values

S_Y = standard deviation of Y

The formula to find r is:

$$r = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{[N\sum X^2 - (\sum X)^2][N\sum Y^2 - (\sum Y)^2]}}$$

After the standard deviation value and the value of r are determined, they are substituted in the formula to give the standard error of the estimate, S_{yx}. Let's see how to calculate the standard error of the estimate from the data in figure 4–30.

This data, together with their sums, are shown in the table. This table shows ΣY = 153.5, ΣY2 = 2726.25, and N = 10. By substituting these values in the formula for the standard deviation for Y, we get

$$S_Y = \sqrt{\frac{\sum Y^2 - \frac{(\sum Y)^2}{N}}{N-1}} = \sqrt{\frac{2726.25 - \frac{153.5^2}{10}}{10-1}} = \sqrt{\frac{2726.25 - 2356.225}{9}} = \sqrt{41.1139} = 6.4120$$

We also solve for r:

$$r = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{[N\sum X^2 - (\sum X)^2][N\sum Y^2 - (\sum Y)^2]}}$$

$$r = \frac{10(3890.5) - (205)(153.5)}{\sqrt{[(10)(5725) - (205)^2][(10)(2726.25) - (153.5)^2]}}$$

$$r = \frac{38905 - 31467.5}{\sqrt{[57250 - 42025][27262.5 - 23562.25]}}$$

$$r = \frac{7437.5}{\sqrt{(15225)(3700.25)}} = \frac{7437.5}{\sqrt{56336306.25}} = \frac{7437.5}{7505.75} = .9909$$

We then substitute this value for S_y (6.4120) together with the calculated value of r (.9909) for this data into the formula for the standard error of the estimate, S_{yx}.

$$S_{YX} = S_Y\sqrt{1-r^2} = 6.4120\sqrt{1-.9819} = 6.4120 \times .1345 = .8624$$

The standard error of the estimate is .8624. This is the calculated value of one standard error of the estimate. Now we plot the standard error of the estimate much the same way as we plotted standard deviation on a control chart.

Using the general formula $Y' = Y \pm ZS_{yx}$.

Where:

Y' = value on standard error of the estimate line (either above or below)

Y = value derived from line of regression calculation

Z = represents your confidence interval

S_{yx} = calculated standard error of the estimate

The standard error of the estimate measures the dispersion of the Y values. Since the Y values are read in the vertical direction, the standard error of the estimate is read in the vertical direction. The standard error of the estimate is measured from each side of the line of regression. Therefore, the standard error of the estimate can be drawn by adding and subtracting its value from a value of the line of regression. For example, the easiest Y values to which we can subtract S_{yx} are the Y values used to plot the line of regression, namely Y = 20.04 when X = 30, and Y = 7.79 when X = 5. Using the Y values from the line of regression calculations (X=5, Y=7.79) and (X=30,Y=20.04), and the calculated standard error of the estimate of .8624, we will calculate the four points necessary to plot the upper and lower standard error of the estimate lines. We will select a Z value of 3 to represent three standard errors of the estimate.

The two points necessary to plot the upper standard error of the estimate line are:

First point:

$Y^{above} = 7.79 + (3)\ (.8624)$

$Y^{above} = 7.79 + 2.5872$

$Y^{above} = 10.4$

Second point:

$Y^{above} = 20.04 + (3)\ (.8624)$

$Y^{above} = 20.04 + 2.5872$

$Y^{above} = 22.6$

We draw a line between the points where X=5 and Y=10.4 and where X=30 and Y=22.6. This line is the upper (above) standard error of the estimate. We also included the line of regression for reference (fig. 4–32).

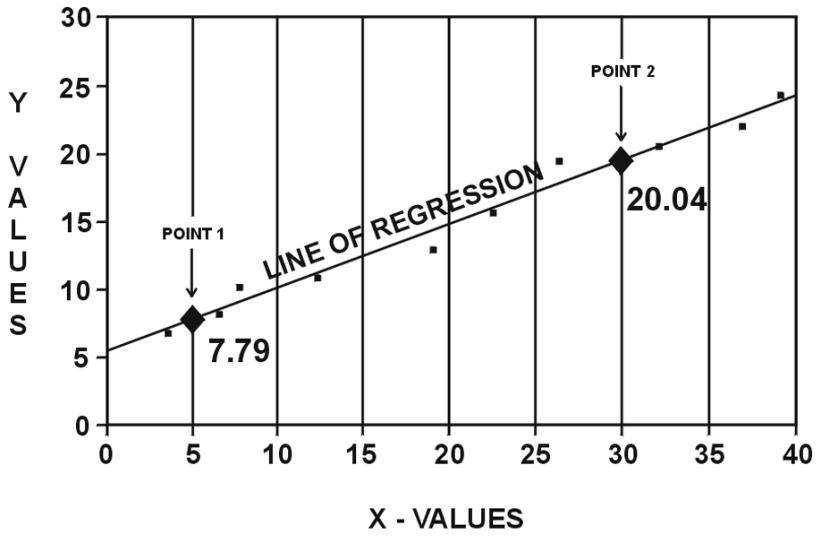


Figure 4–32. Plotting the line of regression.

The two points necessary to plot the lower standard error of the estimate line are:

First point:

$$Y^{below} = 7.79 - (3)\ (.8624)$$

$$Y^{below} = 7.79 - 2.5872$$

$$Y^{below} = 5.2$$

Second point:

$$Y^{below} = 20.04 - (3)\ (.8624)$$

$$Y^{below} = 20.04 - 2.5872$$

$$Y^{below} = 17.4$$

We draw a line between the points where X=5 and Y=5.2 and where X=30 and Y=17.4. This line is the lower (below) standard error of the estimate. Again, we included the line of regression for reference (fig. 4–33).

NOTE: We rounded off the value of the points at the charts in figures 4–32 and 4–33 to the nearest tenth because accuracy is not critical at this point.

The standard error of the estimate is represented by two parallel lines, one on each side of the line of regression. Between these two lines lie approximately up to 99 percent of the tallies if you use up to three standard errors of the estimate. In a similar manner, many other lines can be drawn to mark off any other number of standard errors or confidence intervals as desired. In our example, we could have used ±0.86 (1×0.86) for one standard error of the estimate or ±1.72 (2×0.86) for two standard errors of the estimate.

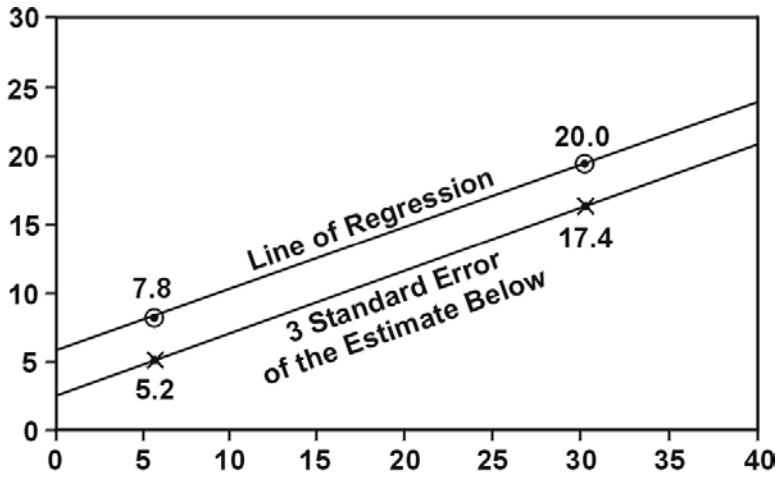


Figure 4–33. Standard error of the estimate below the line of regression.

440. Predicting the trend

We mentioned before that the line of regression and standard error of the estimate are used for prediction. If the relationship between the two series of values isn't perfect, then the predicted values won't necessarily be the same as the actual values (that is, lie on the line of regression). This is because of the scatter or variation about the line of regression. If the scatter is measured (by standard error of the estimate), the variation can be allowed for, and a range can be established within which a given proportion of the values will fall.

Go back to the table in figure 4–30. Let's say that the X values stand for the number of reported minor aircraft discrepancies and the Y values stand for the number of hours it takes to repair the discrepancies. The data in the table then represent the number of minor discrepancies gathered per occurrence (in this case, the occurrence is an aircraft). For the purpose of this lesson, let's define a minor discrepancy as a condition of an aircraft that will take one hour or less to repair (such as replacing a missing screw). Suppose you have just received notification of 20 minor discrepancies on an aircraft. Using the regression analysis chart you have developed with previous data, you read up from 20 on the X-axis until you come to the line of regression.

Then you read off the Y-axis how many hours you can expect to repair those 20 discrepancies. In this case, you can provide an estimated time of 15 hours to complete all jobs (fig. 4–34).

By using the standard error of the estimate, you can get a range or an interval of how many hours you predict are needed to fix an aircraft. Three standard errors of the estimate yield a confidence interval of 99 percent. Thus, you can predict that 99 times out of 100 you will need a certain number of hours to fix an aircraft.

Let's look at the 20 minor discrepancies again. This time, however, we want to use the standard error of the estimate (fig. 4–35) rather than the line of regression to predict how many hours are required to fix an aircraft. We read up from 20 on the X-axis until we come to the lower line for standard error of the estimate. We then read the minimum fix hours we can expect from the Y-axis. We can expect to spend a minimum of 12.5 hours. Similarly, we read up from 20 on the X-axis until we come to the upper line for standard error of the estimate. We then read the maximum repair hours we can expect from the Y-axis. We can expect a maximum of 17.5 hours. In conclusion, for 20 minor discrepancies on an aircraft, 99 times out of 100, we can expect between 12.5 and 17.5 hours to complete all repair on the aircraft. The standard error of the estimate provides a better approximation or estimate than the line of regression.

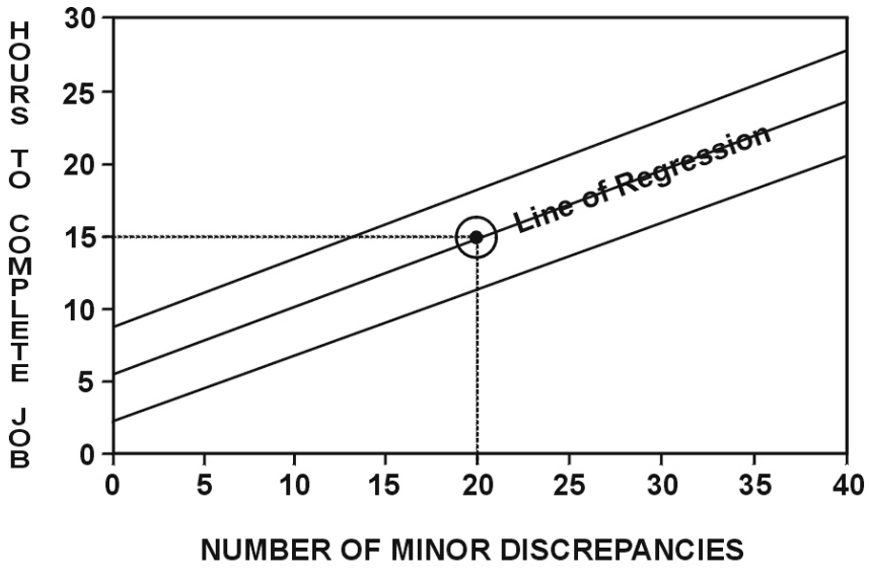


Figure 4–34. Estimated number of hours using line of regression.

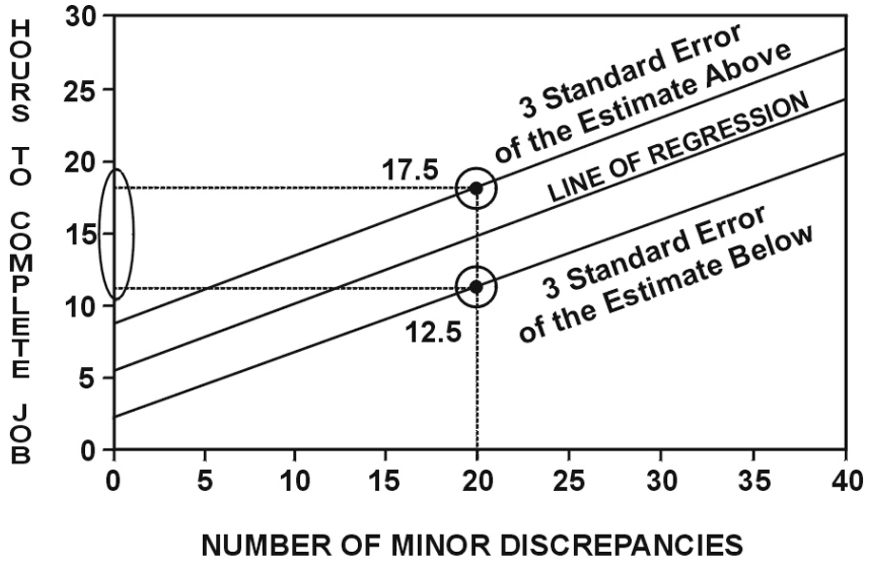


Figure 4–35. Estimated time using standard error of estimate.

Self-Test Questions

After you complete these questions, you may check your answers at the end of the unit.

438. Computing the line of regression

1. What does the line of regression indicate?

2. What technique is used to graph a line of regression?

3. In the formula Y = a + bX, what do a and b represent?

4. Compute the value of the regression line point for X = 10, b = –0.39, and a = 11.5.

439. Computing the standard error of the estimate

1. What does standard error of the estimate measure?

2. Approximately what percent of the tallies fall between the two lines representing one standard error of the estimate on each side of the line of regression?

3. Compute the Y-values representing the points above and below the line of regression using two standard errors of the estimate when S_Y = 1.2, r = 0.8, and the regression line Y value = 14.2.

440. Predicting the trend

1. Approximately what confidence interval can be assigned when three standard errors of the estimate are used for predictions?

2. How can you increase the confidence of a prediction by using the standard error of the estimate?

3. If Y = 20 for a given X value on a regression line, what interval would you predict for future Y values at the same X point if S_{YX} = 2? Use 2 S_{YX} values to make the prediction for a 95 percent confidence interval.

4. When X = 15 on a regression line, what interval can you expect at 3 S_{YX} when S_{YX} = 2.5?

4–5. Probability

Will it rain today? Will my car battery make it another winter? Will this equipment have a major failure within the next 6 months? These should be common questions to you, the person, and to you, the analyst. What answers can you give to these questions? Can you take a definite “yes” or “no” stand on these issues? More than likely, your answer would be something like, “There is a 50 percent chance of rain.” “My battery should last through the winter.” “Chances are good that this piece of equipment will have a major failure within the next six months.”

Each time you answer questions like those, you are using probabilities to infer something. That is a part of inferential statistics. Probabilities have been stated in layman's terms and computed mentally from personal experience, yet these probabilities differ only slightly from those calculated from masses of data.

Before we discuss the practical application of probability, we need to establish a sound basic knowledge of the rules and laws of probabilities. So, we begin the section with a discussion of the basics of probabilities. Then we tell you about classical and frequency probabilities, and we finish up by going over the multiplication and addition rules of probability. Although some of the concepts are theoretical in nature, every attempt is made to give you practical examples of each definition, rule, or law.

441. Probability

Throughout our discussion of probability, you'll be introduced to terms with which you may or may not already be familiar. To understand probability, it's essential that you know these important terms and their definitions.

Mutually exclusive events

An event is an occurrence of something. If something happens or can happen, it's an event. The roll of a die (singular for dice) or drawing a card is an event that can happen. If one event occurring prohibits another event from occurring, the events are called mutually exclusive events.

A maintenance situation that illustrates a mutually exclusive event is a piece of equipment that can be either operable or inoperable. If the equipment is inoperable, it cannot be operable, so the two events are mutually exclusive. This concept can apply to more than two events. For instance, an end item can be red, amber, or green. That is, it is either out of commission (red), operable but is not completely free from defects (amber), or in good shape and capable of operating at maximum design capacity (green). Only one of these conditions can exist at a time on the same piece of equipment. So, you can say that the events red, amber, and green, as defined above, are mutually exclusive.

Independent events

Sometimes the occurrence of one event does not affect the occurrence or likelihood of occurrence of another event. These events, which are in a way not concerned with each other, are known as independent events.

Independent events often seem to exist in maintenance. We say "seem to exist" because they exist as far as we know. There is always a possibility that any one event may have influenced the occurrence or probability of occurrence of another event. One such example is in the coding of maintenance forms. The probability of one work center making errors on 10 percent of its forms will not influence the probability of another work center making errors on 10 percent of its forms. Nor will the probability of one radio channel failing affect the probability of another channel failing. However, if the failure of one channel increases the use of another channel, and its use is directly related to channel failures, the events are not entirely independent. Independent events are defined as those events whose occurrence or probability of occurrence are not affected by each other, as far as we know.

Probability range

Throughout our discussion, we have spoken of probability without fully explaining the term. You know that probability is associated with chance. That is, the chance of something happening (an event) is the probability of that occurrence. What you may not know is that probability can be represented mathematically.

Probability can be expressed only in positive numerical digits between (and including) the numbers 0 and 1—the probability range. Zero probability is the least and one is the most that a probability can be. It is rare for anything to have a probability of either 0 or 1. These probabilities denote absolute

certainty that an event will (1) or will not (0) occur. Is anything that certain? Rather than stating you are certain, often it is advisable to say that an event approaches these probabilities.

You usually describe probabilities by a decimal number between 0 and 1. For example, the probability of obtaining "heads" in one toss of a coin is expressed as 0.5. You can also express this probability in terms other than a decimal. Equivalent expressions of the probability of 0.5 are 1/2 or 50 percent. The majority of the time, however, because of the necessities of calculators, we express probabilities as a decimal figure.

Discrete and continuous variables

Later in this unit we discuss probability distributions that use either discrete data or continuous data. Discrete data are variables that vary only as whole numbers. For instance, if you roll a die, you can expect only whole numbers to show up (e.g., 1, 2, 3, 4, 5, or 6). You could state probabilities about each of these numbers appearing. On the other hand, numbers like 4.3, 2 ½, and so forth are not possible when rolling a die and could not be considered as part of the distribution because they do not vary discretely. Discrete data are countable, such as people, failures, cars, airplanes, radar sets, and so forth.

Continuous data are variables that can be measured in infinite amounts between two given points. Examples are man-hours, speed, time, distance, rates, or almost any data that can be expressed as a ratio. In many instances you can convert discrete data to continuous data by making it a ratio or rate. For instance, although failures are discrete data and end items are considered discrete, failures per end item are considered continuous data because infinitely small rates of failures per end item can exist.

442. Computing classical and frequency probabilities

There are, basically, two definitions of probability in terms of how probability is determined. These are the classical definitions and the frequency definitions. The difference between these definitions is subtle but very important.

Classical definition

The classical definition of probability takes the number of desired results for an event and divides it by the total number of results possible for the event to happen to determine the probability of the desired result. Symbolically this is expressed as:

$$P(A) = \frac{a}{a+b}$$

Where:

P(A) = probability of the event.

(**NOTE**: We can assign any variable other than A to suit the event.)

a = number of ways the desired event can occur.

b = number of ways the desired event cannot occur.

The above formula uses a law of probabilities called the law of complementation. As you know, the maximum a probability can be is 1, indicating that the event had to happen (meaning no other event could happen). As it happens, the total of the probabilities of all the possible events to occur within a closed situation equals 1. In other words, if you roll a die, you can obtain a 1, 2, 3, 4, 5, or 6. The probability of each of these events, when added together, must equal 1 if they represent all possible events. You can easily determine that 0.16 2/3 probability for each event multiplied by the six events gives you 1. Thus, you use a 1 when talking in terms of probabilities. You can symbolize a law of complementation for probabilities to compute the probability that an event will *not* occur. This is written as:

$$P(A) + P(\overline{A}) = 1$$

or

$$P(A) = 1 - P(\overline{A}) \text{ and}$$

$$P(\overline{A}) = 1 - P(A)$$

Where:

$P(\overline{A})$ is the complement of P(A). The symbol $\overline{A}$ means "not A" or "other than A."

Such that:

P(A) = the probability that event A can occur.

$P(\overline{A})$ = the probability that event A cannot occur.

Let's use the law of complementation and the classical definition of probability to solve a problem. Remember, the classical definition of probability uses an entire population of possible occurrences to determine the probability of any one occurrence.

To determine the probability of tossing "heads" in one flip of a coin, you must know the number of ways the desired results could happen. Since you know that only one side of a coin has a head on it, then the number of desired results is equal to 1. The number of nonheads, or undesirable results, possible with one toss is easily determined. Since a coin has only two sides, one of which is a head, you can deduce that the number of nonheads (tails) is also 1. To determine the probability of tossing a head in any one toss of a coin, using the classical definition, simply divide the number of desired results possible (heads) by the number of desired results possible (heads) plus the number of undesired results (tails). So:

$$P = \frac{a}{a+b} = \frac{1}{1+1} = \frac{1}{2} = 0.5$$

Let's use this definition to determine another probability. It might be advantageous to know what chance you have (the probability) of rolling a specific number on a die. If, for instance, you wanted to roll a 3, what are your chances of doing this?

$$P(A) = \frac{a}{a+b}$$

Where:

P(A) = desired probability of rolling a 3.

a = the number of ways the desired event can occur.

b = the number of ways the desired event cannot occur.

So:

$$P(A) = \frac{1}{1+5} = \frac{1}{6} = 0.167$$

Therefore, the probability of rolling a 3 equals 0.167.

The classical definition, as stated earlier, requires that you know the total number of possibilities of all possible occurrences. That is, it works with entire populations. It's practically impossible to use this definition to determine the probability of a component being defective because you could not possibly know in how many ways it could be bad or in how many ways it could be good. Nor could you know how many of this type of component were bad in the past as opposed to how many were good because there are still some in existence, and you do not really know whether or not those are good or bad. Another definition of probability is necessary to handle these types of situations—the frequency definition of probability.

Frequency definition

The frequency definition of probability is much more simply stated than the classical definition. The frequency definition, symbolically stated is:

$$P(A) = \frac{x}{n}$$

Where:

P(A) = the desired probability.

x = the number of occurrences.

n = the number of trials.

This formula means that the probability of occurrence of some event is equal to the number of times the event did happen divided by the number of times the event had a chance to happen. This definition is, in many respects, similar to the classical definition of probability. The difference is one of positive surety. In the classical definition, you're sure of your probability because all possibilities are considered. With the frequency definition, you know only that from your experience with the number of possibilities observed (n), the event occurred so many times the number of possible occurrences (x).

The law of complementation holds true for the frequency definition as well as for the classical definition. Even though it would be rare for you to know the total possible number of events that could occur in a maintenance situation, you still can be sure that, if all the events were considered, the sum of all the probabilities would equal 1.

This definition of probability using frequencies of occurrence is itself subject to certain probabilities of being correct, depending on the size of the sample involved in the determination of the probability. Even though there is some chance involved, using frequencies of occurrence is the best method available to define probabilities for maintenance data. Rarely are you able to use entire populations in your work.

For example, you could not know all of the defects that could possibly occur to a piece of equipment because an infinite number of defects could occur. Therefore, if you desire to know the probability of a piece of equipment having two defects upon its receipt, you could not know the $\overline{A}$ portion of the classical definition. Your alternative to this problem is to use the frequency definition. If, for example, in the past, 20 pieces of equipment were received on base, and of these, two had defects, you could conclude that:

$$P(A) = \frac{x}{n} = \frac{2}{20} = 0.1$$

The probability of the next piece of equipment received having a defect is 0.1 or 10 percent. You can also apply the law of complementation and say that the probability of the next piece of equipment received not having a defect is:

$$P(\overline{A}) = 1 - P(A)$$

$$P(\overline{A}) = 1 - 0.1$$

$$P(\overline{A}) = 0.9 \text{ or } 90 \text{ percent}$$

In this case, the complement of the probability of having defect P(A) is P($\overline{A}$), which represents the nondefective equipment. You should, of course, see that this type of probability is subject to fluctuation when small samples are used. For instance, if n was 3 and x was 1, the probability would be 0.33 that the next unit would have a defect. If that unit does not have a defect, the probability of the next (fourth) unit having a defect would be 0.25, a difference of 0.08.

When the number of trials becomes large, the probability has less tendency toward fluctuation caused by long chance runs of one event's occurrence. For instance, if $n = 1000$ and $x = 10$, then P(A) = 0.01. You can expect the next piece of equipment to have a defect 1 time out of 100 times. If the next piece of equipment does not have a defect and you recompute P(A), you'll obtain:

$$P = \frac{10}{1001} = 0.00999 = 0.01$$

NOTE: The probability is virtually unchanged because of the large sample size.

You have unknowingly applied probability throughout your career. Capability and reliability rates are forms of probability. These rates reflect the probability that a system will fail or malfunction the next time it is used. A radio capability of 96.9 percent indicates it has a 0.969 probability of not having a failure and a 0.031 chance of having a failure.

443. Computing probability laws of multiplication

No doubt, you'll often want to know the probability that two events will occur together or in succession. These events are classed as independent or dependent events. Independent means the occurrence of one event does not affect the probability of occurrence of the other, whereas the probability of dependent events occurring could be affected by other events.

Law of multiplication for independent events

In symbolic language the law of multiplication for independent events is:

$$P_{(1 \text{ and } 2)} = P_1 \times P_2$$

Simply stated, this law says that the probability of event 1 and event 2 occurring is equal to the probability that event 1 occurs multiplied by the probability that event 2 occurs.

What if in the game with the die, you were in a position to enrich yourself if you threw a 2 and a 5, and you wish to throw them in successive rolls of the die. You can determine the probability of this occurrence. To do so, you apply the law of multiplication for independent events.

To throw a 2 and a 5 in succession, you must roll first a 2 and then a 5. You know that the probability of rolling any given number on a die is .167. Then, the probability of rolling a 2 and a 5 in succession is:

$$P_{(2 \text{ and } 5)} = P_2 \times P_5$$
$$P_{(2 \text{ and } 5)} = (0.167)\ (0.167)$$
$$P_{(2 \text{ and } 5)} = 0.028$$

Thus, you could expect to be lucky 28 times out of 1,000 tries.

Let's apply this rule to maintenance data. Assume two pieces of equipment are independent of each other in respect to their probability of failing. One piece of equipment has a 10 percent chance of failing each time it is turned on. The other has a 0.1 percent chance of failing when it is turned on. What is the probability that a maintenance technician would walk in one morning, turn on both pieces of equipment, and find that neither of them is working? This situation would be:

$$P_{(1 \text{ and } 2)} = P1 \times P2$$
$$P_{(1 \text{ and } 2)} = (0.1)\ (0.001)$$
$$P_{(1 \text{ and } 2)} = 0.0001$$

So you can say that the probability that neither piece of equipment will work when first turned on is 0.0001 or 1 in 10,000.

Law of multiplication for dependent or conditional events

The law of multiplication for dependent events states that the total probability is the product of the probability of one of the events occurring, and the probability of the other occurring, given that the first has already occurred. This law is symbolically stated as:

$$P_{(1 \text{ and } 2)} = P_1 \times P_{(2/1)}$$

Where:

$P_{(1 \text{ and } 2)}$ = probability of both events occurring.

P_1 = probability of event 1.

$P_{(2/1)}$ = probability of event 2, given that event 1 has already occurred.

This law can be illustrated easily with cards. To draw two clubs in succession from a deck of cards, the probability of event 1 occurring is $^{13}/_{52}$, since there are 13 clubs in a deck of 52 cards. The probability of event 2 occurring, given that event 1 has occurred is 12/51, since there are 12 clubs left in a deck of 51. Remember, you are saying that event 1 occurred.

Substituting in the formula and multiplying these two probabilities together, you get:

$$P_{(1 \text{ and } 2)} = P_1 \times P_{(2/1)}$$

$$P_{(1 \text{ and } 2)} = \left(\frac{13}{52}\right)\left(\frac{12}{51}\right)$$

$$P_{(1 \text{ and } 2)} = 0.0588$$

As you can see, the occurrence of event 1 affected the probability of event 2 occurring, making the event dependent.

444. Computing probability laws of addition

The rule of addition deals with the conjunctions *or* and *either*, meaning that you want to know the probability of one event or another occurring. The word *or* can be rather tricky. Grammatically speaking, there are two *ors*.

One *or* is called the *exclusive or*. This *or* can be illustrated by the probability of drawing either a king or an ace from a deck of playing cards. It is obvious in this situation that you cannot draw both a king and an ace at the same time because the two events are mutually exclusive. The exclusive *or* allows the occurrence of one event or the other, but not both. You are interested in the exclusive *or* in relation to the law of addition for mutually exclusive events.

The other *or* is called the *inclusive or*. An example of the inclusive *or* is illustrated in the probability of drawing either an ace or a heart from a deck of playing cards. Obviously, you could draw either an ace or a heart, but you could also draw the ace of hearts. Simply, then, the inclusive *or* allows the possibility of either event or both events happening at the same time. In this situation you use the law of addition when the events are independent.

Law of addition for mutually exclusive events

The law of addition for mutually exclusive events is another simple concept. It simply means that if two (or more) events are mutually exclusive, the probability of occurrence of any of these events is the sum of the probability of occurrence of all the events. Here, you are interested in only *one occurrence*. Symbolically this law is:

$$P_{(1 \text{ or } 2)} = P_1 + P_2$$

A classic example of this probability law that has already been mentioned is drawing a king or an ace from a deck of playing cards. This probability is expressed by:

$$P_{(K \text{ or } A)} = P_K + P_A$$

$$P_{(K \text{ or } A)} = \frac{4}{52} + \frac{4}{52}$$

$$P_{(K \text{ or } A)} = \frac{8}{52}$$

$$P_{(K \text{ or } A)} = 0.1540$$

For a frequency example, recall the section on mutually exclusive events where we discussed the probability of an end item being either red, amber, or green. Assuming that these events are truly mutually exclusive, you can determine the probability of this item of equipment being reported as other than fully operational, without knowing the probability of its being operational. That is, you wish to determine the probability of the equipment being reported red or amber.

If past data indicate that this equipment is amber 22 hours and red 19 hours in a 720-hour month, then:

$$P_{(R \text{ or } A)} = P_R + P_A$$

$$P_{(R \text{ or } A)} = \frac{19}{720} + \frac{22}{720}$$

$$P_{(R \text{ or } A)} = \frac{41}{720}$$

$$P_{(R \text{ or } A)} = 0.057$$

You can expect that something will be wrong with this equipment 5.7 percent of the time.

Law of addition for independent events

There may also come a time when you'll want to know the probability of *either or both events* occurring. You may want to know the probability of getting an ace or a heart in a draw from a deck of cards. In this case, you could draw an ace or a heart or both (the ace of hearts) in a single draw. This fact makes the event independent.

To solve this problem, apply the second law of addition, which states that the probability of the occurrence of either or both of two events is the sum of their individual probabilities, minus the probability of both events occurring at the same time. Symbolically stated:

$$P_{(A \text{ or } B)} = P_A + P_B - P_{(A \text{ and } B)}$$

In the case of drawing an ace (event A) or of drawing a heart (event B), add P_A which considers four aces in the deck, to P_B, which considers the 13 hearts in the deck, and subtract the $P_{(A \text{ and } B)}$ because you have considered the ace of hearts twice. P_A is equal to 4/52 and P_B is equal to 13/52. P_A indicates that you have four aces (including the ace of hearts). P_B also includes the ace of hearts. This points out the reason why you subtract the P (A and B). Solving the problem:

$$P_{(A \text{ or } B)} = P_A + P_B - P_{(A \text{ and } B)}$$

$$P_{(A \text{ or } B)} = \frac{4}{52} + \frac{13}{52} - \frac{1}{52}$$

$$P_{(A \text{ or } B)} = \frac{16}{52} = 0.3077 \text{ or } 30.77\%$$

Therefore, the probability of drawing an ace or a heart from a deck of cards is 30.77 percent.

In the previous example you knew the probability of A and B occurring in a deck of cards was $^1/_{52}$; however, this probability is not always known when you are dealing with typical maintenance data.

To determine the probability of occurrence of A and B simultaneously, simply multiply A times B. Expressing the example symbolically:

$$P_{(A \text{ or } B)} = P_A + P_B - P_{(A \times B)}$$

$$P_{(A \text{ or } B)} = \frac{4}{52} + \frac{13}{52} - \left(\frac{4}{52} \times \frac{13}{52} \right)$$

$$P_{(A \text{ or } B)} = \frac{17}{52} - \left(\frac{52}{2704} \textbf{ or } \frac{1}{52} \right)$$

$$P_{(A \text{ or } B)} = \frac{16}{52} = 0.3077 \text{ or } 30.77\%$$

You may have learned from experience that the probability of completing a periodic inspection is 0.96 (event A). The probability of completing an engine change is 0.90 (event B), and the probability that these two events will occur together is 86.4 percent (0.96 × 0.90) of the time.

Solving the maintenance example then:

$$P_{(A \text{ or } B)} = 0.96 + 0.90 - (0.96 \times 0.90)$$

$$P_{(A \text{ or } B)} = 0.96 + 0.90 - (0.864)$$

$$P_{(A \text{ or } B)} = 0.996 \text{ or } 99.6\%$$

This indicates that there is a 99.6 percent chance that event A or event B or both will be completed within the time limit.

Two good points to keep in mind when dealing with the laws of addition are: (1) they are used to compute probabilities for statements using the conjunctions *either* and *or*, and (2) the laws for mutually exclusive events and independent events are slightly different.

Self-Test Questions

After you complete these questions, you may check your answers at the end of the unit.

441. Probability

1. Define the following terms:

   a. Mutually exclusive events.

   b. Independent event.

   c. Probability range.

   d. Discrete data.

   e. Continuous data.

2. Label each of the following events as either mutually exclusive or independent:

   a. Tossing a head and tossing a tail in one toss of a coin.

   b. Tossing a head in one toss of a coin, and tossing a head on another toss of the coin.

   c. Finding the left front tire flat on your car or finding the right front tire flat on your car.

   d. Rolling a one or a four on one roll of a die.

3. What type of data can be expressed in infinitely small portions?

4. How can an analyst convert discrete data to continuous data?

442. Computing classical and frequency probabilities

1. State the classical definition of probability in symbolic terms.

2. Why can't you usually use the classical definition of probability with maintenance data?

3. State, in symbolic terms, the frequency definition of probability.

4. Over the past 6 months, one work center has submitted 650 like maintenance forms, 24 of which have contained errors. What is the probability that the next form submitted by this work center will be in error?

5. If there are 6 types of errors that occur on forms and one type occurs 10 percent of the time, with what probability will the remainder of the errors occur?

443. Computing probability laws of multiplication

1. Symbolically state the probability laws of multiplication for both independent and dependent events.

2. If work center A has a 0.88 probability of completing the mission, and work center B has a 0.94 probability of completing the mission, what is the probability that both work centers will complete the mission? The events are independent.

3. Out of 1,200 jobs scheduled, 780 were started as scheduled. Of the 780 jobs started as scheduled, only 663 were completed. What is the probability that a job will be started as scheduled (event 1) and then completed (event 2).

444. Computing probability laws of addition

1. Symbolically state the probability law of addition for mutually exclusive events.

2. Symbolically state the probability law of addition for independent events.

3. What is the probability of drawing either a club or a queen in one draw from an ordinary deck of 52 playing cards?

4. Out of 2,000 jobs scheduled, 60 were canceled for maintenance reasons and 20 were canceled for supply reasons. What is the probability that a scheduled job will be canceled either for maintenance or supply reasons?

5. A particular type of equipment has two UHF radios. The probability of UHF radio number 1 functioning properly is 0.80 and the probability of UHF radio number 2 functioning properly is 0.90. Either UHF number 1 or number 2 must function properly to complete the mission. What is the probability that the mission will be completed?

6. Out of 135 jobs completed, 22 were done by E–4s and 88 were done by E–5s. What is the probability that an E–4 or an E–5 will complete the next job? Consider these events mutually exclusive.

Answers to Self-Test Questions

429

1. The degree of relationship between two or more sets of data.
2. A number that tells to what extent two sets of data are related.
3. 0.97 and –0.97 have identical degrees of relationship.
4. A graph or picture showing the relationship between two sets of measures.
5. One pair of X and Y values.
6. (1) d.
   (2) a.
   (3) b.
   (4) c.

430

1. r.
2. The degree of association or linear relationship between two variables.
3. The data must have been selected at random and in pairs from a normal distribution. It must display homogeneity of variance and be from the interval or ratio measurement scale.
4. The calculated r value must exceed the table value to be a significant relationship.
5. r = .85.

431

1. They are de-emphasized because of ranking the data.
2. When an estimate of the degree of correlation is needed.
3. Rho (ρ).
4. The computed value of rho is normally smaller than the computed value of r.
5. Rho (p) = .99.

432

1. Data observations over a consecutive period of time.
2. Secular represents the natural forces acting on the data, seasonal refers to forces caused by calendar time periods, and cyclical refers to forces caused by the nature of the data itself.
3. (1) It must be possible to categorize the data by some time period, preferably by month for maintenance data. This allows the formulation of a time series.
   (2) The time series that represents the data must be at least 24 months in duration. Anything less is difficult to visualize.
   (3) The data, when plotted, must have a reasonable resemblance to a straight line.
4. The arithmetic mean.

433

1. $a = \frac{\Sigma Y}{N}, b = \frac{\Sigma XY}{\Sigma X^2}$, and $Y_c = a + bX$
2. May = 4.6013; September = 3.0645.
3. Nov 2002 = –3, Dec 2002= –1, Jan 2003 =1, and Feb 2003= 3.
4. Nov 2002= –2, Dec 2002= –1, Jan 2003= 0, and Feb 2003= 1.

434

1. If the data does not meet linear trend prerequisites.
2. The second-degree parabola method.

3. (1) Data at each end of the overall series is lost.
   (2) There is no way to determine the exact position of the trend line value for the most recent time period.
   (3) Extrapolation is not possible when using this method.
4. $Y_C = a + bX + cX^{2.}$

435

1. 100 percent.
2. 8⅓.

436

1. 110.7 percent.
2. .0481 or 4.81 percent.
3. .1804.
4. 120.
5. 107.

437

1. A method used to test the significance of a linear trend.
2. The values that are tied are given the average rank they are tied for.
3. Positive indicates an increasing trend. Negative indicates a decreasing trend.
4. Z_S = 1.14 or 74.58 percent—which indicates an insignificant trend—74.48 percent is less than the 95 percent recommended before you investigate, therefore chance causes are not affecting the data.
5. Do not investigate; 74.58 is less than 95.

438

1. The trend of the data relationships.
2. The least-squares method.
3. a = the point where the line crosses the Y axis, b = the slope of the line.
4. Y = 7.6.

439

1. The dispersion about the line of regression.
2. 68 percent.
3. 15.64 and 12.76.

440

1. 99 percent.
2. By increasing the number of standard error of the estimates used.
3. 16 – 24.
4. 7.5 – 22.5.

441

1. a. When one event occurring prohibits another from occurring.
   b. The occurrence of one event that does not affect the occurrence or likelihood of another event occurring.
   c. 0–1.
   d. Data that vary only as whole numbers.
   e. Data that can be measured in infinite amounts.
2. a. Mutually exclusive.
   b. Independent.
   c. Independent.
   d. Mutually exclusive.

3. Continuous data.
4. By making the data a ratio or rate.

442

1. $P(A) = \dfrac{a}{a+b}$
2. Because the classical definition requires you to know all the possibilities of occurrence, an impossible task in maintenance.
3. $P(A) = \dfrac{x}{n}$
4. $P(A) = \dfrac{24}{650} = 0.037$
5. 0.90 or 90 percent.

443

1. a. The law of multiplication for independent events = $P_{(1 \text{ and } 2)} = P_1 \times P_2$.
   b. The law of multiplication for dependent events = $P_{(1 \text{ and } 2)} = P_1 \times P_{(2/1)}$.
2. P_1 and $P_2 = P_1 \times P_2 = (0.88)(0.94) = 0.827 = 0.83$.
3. $P_{(1 \text{and } 2)} = P_1 \times P_{(2/1)} = \dfrac{780}{1200} \times \dfrac{663}{780} = 0.65 \times 0.85 = 0.5525 = 55.25\%$

444

1. $P_{(1 \text{ or } 2)} = P_1 + P_2$.
2. $P_{(A \text{ or } B)} = P_A + P_B - P_{(A \text{ and } B)}$.
3. 30.77 percent.
4. $P_{(1 \text{ or } 2)} = P_1(0.03) + P_2(0.01) = 0.04$.
5. $P_{(A \text{ or } B)} = 0.80 + 0.90 - 0.72 = 0.98$.
6. $P_{(1 \text{ or } 2)} = P_1 + P_2 = \dfrac{22}{135} + \dfrac{88}{135} = \dfrac{110}{135} = 0.8148$

Complete the unit review exercises before going to the next unit.

Unit Review Exercises

Note to Student: Consider all choices carefully, select the *best* answer to each question, and *circle* the corresponding letter. When you have completed all unit review exercises, transfer your answers to the Field-Scoring Answer Sheet.

Do not return your answer sheet to AFCDA.

68. (429) When performing correlation analysis, as values in one set of measure increase *and* the corresponding or paired values in the other set also increase, the relationship is considered to be
    a. positive.
    b. negative.
    c. a coefficient of correlation.
    d. a low degree of scattered data.

69. (430) When the calculated value of Pearson's r *exceeds* the coefficient of correlation from the table of critical values, what does that indicate?
    a. A correlation exists and it is significant.
    b. A correlation exists but it is not significant.
    c. No significant correlation exists between the data sets.
    d. There is no significant difference between the data sets.

70. (431) What is the *full* range of Spearman's rank correlation coefficient?
    a. –1 to +1.
    b. –1 to 0.
    c. 0 to +1.
    d. +1 to +2.

71. (432) In trend analysis, a set of data observations made at consecutive time periods is called a
    a. time series.
    b. secular series.
    c. monthly series.
    d. cyclical observation.

72. (432) In trend analysis, the *most* general, common type of variation in a set of data over a *long* period of time can be described by a
    a. time series.
    b. seasonal trend variation.
    c. cyclical trend variation.
    d. secular trend variation.

73. (432) For a time series-type of trend analysis, what is the recommended *minimum* period of time for the amount of data required?
    a. 6 months.
    b. 12 months.
    c. 18 months.
    d. 24 months.

74. (432) What statement *best* describes the cyclical variation type of trend analysis?
    a. Usually caused by intervals shorter than a year.
    b. Usually caused by the nature of the data itself.
    c. The most general and common type of variation.
    d. The most accurate method used for prediction.

75. (432) What is the *minimum* time period required to use linear trend analysis techniques?
    a. 6 months.
    b. 12 months.
    c. 18 months.
    d. 24 months.

76. (432) What month or months is/are *not* used in plotting a semi-average trend analysis line?
    a. Middle.
    b. End only.
    c. Beginning only.
    d. Beginning and ending.

77. (433) Whenever data tend to form a straight line, what is the *most* popular method for computing the secular trend analysis of a time series?
    a. Line of regression.
    b. Time series analysis.
    c. Least-squares method.
    d. Parabolic trend analysis.

78. (433) When computing a trend line using the least-squares analysis method, what does "b" represent in the formula?
    a. Slope of the trend line.
    b. Computed average.
    c. Assigned variable.
    d. A constant.

79. (433) Using the formula $Y_C = a + bX$, compute Y_C where a = 2.545, b = 0.1509 and X = –5.
    a. 1.791.
    b. –1.791.
    c. 2.791.
    d. –2.791.

80. (433) Given that $Y_C = a + bX$, $a = \Sigma Y/N$, $b = \Sigma XY/\Sigma X^2$, N= 24, $\Sigma Y = 1800$, $\Sigma XY = 4700$, and $\Sigma X^2 = 12{,}100$, what is Y_C for X = 23?
    a. 66.07.
    b. 74.61.
    c. 75.39.
    d. 83.93.

81. (434) Which plots are used to draw the moving-average in a nonlinear trend line?
    a. Each plot in the series.
    b. The two extreme plots in the series.
    c. The center plots for each time span.
    d. The plots closest to the overall series average.

82. (434) What is considered a *disadvantage* when using the moving average trend analysis method?
    a. Extrapolation is not possible.
    b. Seasonal variations cannot be smoothed.
    c. Cyclical variations will continue to fluctuate.
    d. The data cannot be used for any other time series.

83. (435) To analyze *seasonal* trends, you would use the percent-of-yearly-total method to establish
    a. an index of seasonal variations.
    b. an extrapolation of seasonal data values.
    c. the probability of seasonal trend variations.
    d. the line of regression for seasonal data values.

84. (435) When using a seasonal index, the 8⅓ percent ($^{100}/_{12}$) centerline serves as a reference point to indicate the months
    a. within the prediction interval.
    b. above or below the overall average.
    c. increasing at a steady rate of change.
    d. between the upper and lower control limits.

85. (436) The use of statistical methods in the forecasting of secular trends analysis consists *primarily* of
    a. determining variance.
    b. using scatter diagrams.
    c. measuring events in the past.
    d. calculating the probability of occurrence.

86. (436) If 140 *seasonal* deviations occurred in 2015 and 150 occurred in 2016, how many deviations would you predict for 2017?
    a. 130.
    b. 135.
    c. 155.
    d. 160.

87. (437) The standard deviation of "S" is dependent on the
    a. mean.
    b. population.
    c. sample size.
    d. computed value of "S."

88. (437) What is the *recommended minimum* probability of significance of a linear trend before further investigation is warranted?
    a. 85 percent.
    b. 90 percent.
    c. 95 percent.
    d. 99 percent.

89. (438) In regression analysis, what does a line of regression indicate?
    a. Probabilities of the data.
    b. Standard errors of the data.
    c. Trend of the data relationships.
    d. Coefficient of correlation of the data relationships.

90. (438) If $b = \Sigma XY/\Sigma X^2$ and $a = \Sigma Y/N$, determine the values of a and b when $\Sigma XY = 1800$, $\Sigma X^2 = 2600$, $\Sigma Y = 200$, and $N = 30$.
    a. a = 1.44; b = 6.67.
    b. a = 6.67; b = 1.44.
    c. a = .69; b = 1.44.
    d. a = 6.67; b = .69.

91. (439) Compute the standard error of the estimate limit of one S_{YX} when $S_Y = 2$, r = .1, and the regression line Y value is 10.
    a. 8.02 and 11.98.
    b. 8.02 and 10.00.
    c. 10.00 and 11.98.
    d. 10.20 and 9.98.

92. (440) When plotting a line of regression, you allow for variation of the predicted values by
    a. using probability analysis.
    b. calculating the standard error of the estimate.
    c. determining the proportion of predicted values to actual values.
    d. establishing a range within which a given proportion of the values will fall.

93. (440) What would be the interval when X = 25 and $S_{YX} = 5$ at three standard errors of the estimate?
    a. 5–40.
    b. 10–40.
    c. 15–25.
    d. 20–40.

94. (441) In probability statistics, what type of data are numbers of personnel, failures, cars, airplanes, and radar sets?
    a. Continuous.
    b. Discrete.
    c. Interval.
    d. Ratio.

95. (442) Using the frequency definition for probability, compute the probability of the next work order being in error if 40 errors are found in 400 work orders. Formula:
    $$P(A) = \frac{X}{n}$$
    a. 0.08.
    b. 0.09.
    c. 0.10.
    d. 0.11.

96. (443) If one piece of equipment in a shop fails 5 percent of the time while it is turned on, and a separate piece of equipment fails 10 percent of the time while it is turned on, what is the probability they will both fail when turned on at the same time? Probability formula for law of addition for independent events:
    $P_{(1 \text{ and } 2)} = P_1 \times P_2$.
    a. 0.5 percent.
    b. 1.0 percent.
    c. 1.5 percent.
    d. 50.0 percent.

97. (444) When you need to know the probability of two independent events occurring at the *same time*, you determine the answer by
    a. multiplying the two probabilities.
    b. subtracting the two probabilities.
    c. dividing the two probabilities.
    d. adding the two probabilities.

98. (444) To solve this problem, apply the law of addition for independent events. Of 100 pieces of equipment in your shop, you had 20 scheduled maintenance and 10 unscheduled maintenance performed. What is the probability your next piece of equipment will need either scheduled, unscheduled, or both types of maintenance?
    a. 28 percent.
    b. 32 percent.
    c. 40 percent.
    d. 60 percent.

Glossary

Terms

alternative hypothesis—An assumption about a population parameter that will reject the null hypothesis

bias sample—A sample prejudiced toward one view (intentional or unintentional).

coefficient—A single number representing the amount of change or effect in a process.

correlation—The relationship between variables.

cumulative frequency distribution—A distribution that accumulates its values from class to class.

cyclical variation—The natural flow of data with no outside influences.

degrees of freedom—The number of times data change in a series, N–1.

deviate—A value which differs from the average or normal.

discrete data—Whole numbers only.

extrapolation—Extending a linear trend line for the purpose of making predictions.

frequency polygon—A graph connecting the midpoints of several frequency classes.

harmonic mean—An average of rates (e.g., rate of time or rate of speed).

histogram—A graph showing the ranges of classes within a distribution.

hypothesis—An assumption to explain an observation or occurrence that can be tested for further investigation.

independent events—The occurrence of one event does not affect the occurrence or likelihood of another event.

mean—The numerical average of a series.

median—The positional average of a series.

mode—The most frequent value of a series.

moving average—The sequential averages of subgroups in a series.

mutually exclusive events—One event occurring prohibits another event from occurring.

null hypothesis—A statement about a population or population parameter that is assumed to be true.

parameter—Measures that describe data populations, normally symbolized by Greek letters.

probability range—a measure or estimate of the degree of confidence one may have in the occurrence of an event, measured on a scale from zero (least) to one (most).

regression—Measures that validate the relationship between variables.

sampling distribution—The average distribution of sample values, normal or typical reference for comparison.

secular variation—Trend due to assignable cause.

seasonal variation—Changes in data caused by elements of calendar time.

significance level—The value of α that gives the probability of committing a type I error.

standard deviation—The average deviation of data values from their mean.

standard error—The measurement of difference between a sample's values and its population's values.

stratified sampling—Subgrouping data before sampling the subgroups.

systematic sampling—Taking every Nth item in a series.

symmetrical distribution—Data that has equal quantities on both sides of the mean.

time series—Data measured across a period of time.

trend—The secular variation in the flow of data.

type I error—An **e**rror that occurs when a true null hypothesis is rejected.

type II error—An error that occurs when a false null hypothesis is not rejected.

variance—The average deviation of the square of a variable about the square of the mean, standard deviation squared.

variability—Dispersion or scatter of values within a distribution.

Symbols

α	alpha; significance level, used to determine confidence level of probability
β	beta
$\mathbf{X^2}$	chi-square test
$\mathbf{Y_c}$	computed trend value
df	degrees of freedom
r	denotes Pearson's coefficient of correlation
D	difference
$\neq$	does not equal
$\geq$	equal to or greater than
$\leq$	equal to or less than
=	equals
$S_{\bar{X}}$	estimate of the standard error of the mean

f	frequency
>	greater than
l	identifies class lower limits in a frequency distribution
X	individual value
∞	infinity
<	less than
$\overline{X}$	mean
$\overline{\overline{X}}$	mean of mean
md	median
mo	mode
H_0	null hypothesis
Z	number of standard deviations on a normal curve table
n	number of values in a sample (sample size)
N	number of values in a series
%	percent
μ	population average or mean; Greek letter mu
$\sqrt{}$	radical (square root)
R	range of data elements in a series
$\mathbf{R_Y}$	Rank of Y
ρ	rho; denotes Spearman's coefficient of correlation
Σ	shows the summation of a series of values; Greek capital letter sigma
D^2	Square of the difference
σ	standard deviation of a population; Greek lower-case letter sigma
s	standard deviation of a sample
$\mathbf{s^2}$	standard deviation squared
$\sigma_{\overline{X}}$	standard error of the mean
$\mathbf{R_X}$	symbol for rank of X
(–)	when a bar is placed above a symbol, this indicates an average of the values represented by the symbol (e.g., $\overline{X}$ = mean of Xs)

Abbreviations and Acronyms

AFTO	Air Force technical order
CL	centerline
df	degrees of freedom
IMDS	Integrated Maintenance Data System
IQ	intelligence quotient
JDD	job data documentation
LCL	lower control limit
MC	mission capable
MTBF	mean time between failures
NMCS	not mission capable supply
PEMDAS	parenthesis, exponents, multiplication, division, addition, subtraction
QLP	query language processor
SE	support equipment
TDY	temporary duty
TO	technical order
UCL	upper control limit

Student Notes

AFSC 2R051
2R051 03 1705
Edit Code 07