Statistics

Have you ever noticed how much of the information you receive comes to you through numbers? Particularly since the coming of age of electronic computers, more and more kinds of information are being coded, processed, and presented numerically. On any given day, we can expect to see numerical presentations of weather information, the stock market, political polls, business transactions, census data, government operations, and many other types of data.

In most instances, the numerical information in its original form would be difficult to interpret. For this reason, the information is usually organized, summarized, and presented to us in a form that can be more readily interpreted. Frequently this is accomplished by reducing the numerical data to a table or graph or by reporting one number, such as the average, to represent an entire set of numbers. The process by which numerical data are collected and eventually presented in a usable and understandable form is an important part of the mathematical science of *statistics*.

Statistics is important not only for communication; it also provides a basis for decision making. The government makes extensive use of statistics in estimating its budget needs and setting its tax rates. Statistics enables manufacturers to compare production processes when they seek to improve their products or increase their profits. Store managers may rely upon statistical analyses to determine which items they should stock. Scientists employ statistics in comparing the effects of critical variables upon their experiments. Insurance companies rise and fall on the accuracy of their statistical predictions. Engineers base the design of highways and bridges upon statistical studies of materials and traffic. School officials may modify their curricula on the basis of statistical analyses of student achievements and needs. The list of such decision-making uses of statistics is almost endless.

Collection of Data

The science of statistics involves a variety of tasks. Even the seemingly simple business of collecting numerical data requires careful study. Obviously the conclusions of a statistical study can be no more reliable than the figures upon which they are based. The statistician must be sure that the data collected are accurate, relevant to the problem being studied, and representative of the problem. Invalid conclusions drawn from statistical evidence are often due to inadequacies in the data collected. The matter of data collection will not be discussed in detail, but its importance is so great that we should be aware of some of the problems involved.

First, the *population* to be studied must be well defined. What do we mean by "population" here? To the statistician, it may consist of a set of cities, automobiles, books, or even scientific experiments. In fact, the population for a statistical study might be any set of objects having a common characteristic to which a number might be assigned. Of course, the population selected must supply the appropriate numerical data for the problem being studied. If the population of a statistical study is not well defined or is not representative of the problem being studied, the results of the study will be difficult to interpret or apply. For example, surveys of the voting preferences of high-school students would be of questionable value in predicting the results of a national election, since few high-school students can vote.

Once the population has been identified, the particular characteristics to be studied must be represented numerically. Sometimes the numerical data are already available in recorded form. For example, if you wanted to study the rainfall in your city over the past year, you could probably obtain the needed data from your local weather bureau. In this case, the population might be defined as the set of days in the year.

Sometimes the numerical data may be obtained by a simple counting process. You might be interested, for example, in a study of the books in a school library. This project would involve counting the volumes devoted to each of several subjects.

More often, the data for a statistical study are obtained by measuring some common characteristic of the population being studied. If the population were a set of scientific experiments, the scientist might be concerned with such characteristics as time, temperature, volume, and mass. In each case, the scientist would need to use a suitable measuring instrument to assign a number to the characteristic.

In some cases, no measuring instrument is available and the investigator must create a measuring device. A teacher, for example, creates examinations. Each classroom test is an instrument by which student achievement can be measured. Test grades are the numbers assigned to that measurement. As with any other measuring instrument, accuracy is a prime consideration. The investigator would need to know how well the number assigned represents the true value of the characteristic being measured—in this case, student achievement.

The ultimate value of a statistical study depends to a large extent upon the quality of the measuring instrument that is employed. For this reason, the construction and evaluation of such an instrument is often a critical task in a statistical study

Sometimes it is possible to obtain numerical data about each member of a population that is being studied. When this is true, the data are completely representative of the population, and the task of the statistician is to describe the numerical data obtained. This branch of statistics is called *descriptive statistics*.

Sampling a Population

Often it is necessary or practical to collect data only for a *sample* of the population and to make *statistical inferences* about the population itself. An inference is a conclusion about the unknown based upon something that is known. A statistical inference, therefore, is one based upon statistical data. When data are available only for a sample, the sample represents the known and the population the unknown. Any subject of a population would constitute a sample, but statistical inferences are valid only when the sample is representative of the population. Many techniques are employed by statisticians to insure that the samples they select are representative.

When each member of the population has an equal chance of being chosen, we have what is called a *random sample*. This is usually assumed to be representative of the population. In some special problems, the statistician uses a *stratified sample,* which insures that specific segments of

the study population are represented in the sample. The process of identifying a representative sample is a critical task in many statistical studies. In the examples cited in this article, we will assume that the samples used are representative of the populations from which they have been selected.

To summarize the discussion of collecting data, let us consider the following example: Suppose that the population being studied is the set of students in a given grade. If each of these students was assigned to an English class by a random process, the English class would constitute a representative sample of the population. If you were to measure the height, weight, or age of each student in the class, or record each student's score on a particular test, or count the number of people in his or her family, you would obtain a set of numbers. These numbers could then become the data for a statistical study. The data could be used to describe the English class (the sample). They could also be used to make *estimates* or inferences about the total set of students in the grade (the population).

Organizing the Data

Once data have been collected, they must be arranged in some systematic order before a useful interpretation can be made or conclusions drawn. Sometimes a simple table or graph can be quite helpful as a first step toward the statistical analysis of numerical data.

The numerical data presented in Table I are scores obtained by 35 students from one sixth-grade class, Class 6-1, in a vocabulary test. Each score represents the number of test questions that have been answered correctly by a given student. In this case, the word "score" is used in its usual sense. However, regardless of the nature of the numerical data, statisticians often use the term *raw score* to indicate the individual numbers obtained as a basis for a statistical study.

It is difficult to make any useful interpretation of the data in Table I. The simplest way of organizing these data would be to arrange the scores in numerical order. It is common to record only the different raw scores (in this case, 16, 18, 14, and so on) and to note the *frequency* with which each score occurs. Table II is a frequency table prepared from the data in Table I.

Even a cursory examination of Table II permits some elementary interpretation of the data. We can easily observe the highest and lowest scores (20 and 12) and the most frequent scores (15, 16, and 17). We can even begin to have some feeling for the way the scores seem to cluster about a central point—in this case, the score 16.

Further clarification of the data may be obtained by translating Table II into a graphical form. Figure 1 is a common type of *frequency graph* used to present frequency distributions. The numbers below the horizontal line represent scores; the numbers at the left represent the frequency distribution—that is, the number of times each score occurs.

The *frequency polygon* shown in Figure 2 is based on the same idea. In this case, we imagine lines perpendicular to the scores on the bottom line and lines perpendicular to the frequencies at the left. A point indicates the intersection of a score line and a frequency line. Thus we have a

point where the score 12 and the frequency number 2 meet, and a point where the score 16 and the frequency number 9 meet. The points are connected by straight lines. Note that to "complete the appearance" of the polygon, a zero-frequency-point is assigned to the value one unit lower than the lowest score and to the value one unit higher than the highest score.

An examination of these two graphs will reveal exactly the same information available in Table II. The graphical form often is easier to interpret than the tabular form. In particular, the tendency of scores to cluster around a central point becomes more apparent when the data are presented pictorially.

Grouping Raw Scores

The frequency distribution for raw scores made by all 207 students from all sixth grade classes in the school is shown in Table III. In many statistical studies, particularly when the range of scores is great, it becomes cumbersome to work with all the individual scores. In these cases, it is common to condense the data by grouping the raw scores into class intervals. In studying the scores on the vocabulary test, instead of considering each score individually, we combine a number of adjacent scores to form an interval. Thus we would combine the individual scores 20, 21, and 22 to form the interval 20–22. Intervals are treated in much the same way as we would treat individual raw scores.

A grouped frequency table, based on intervals, is shown in Table IV. It presents scores made by 207 students from all the sixth-grade classes.

When the number of possible individual scores is large, the grouping of data into appropriate intervals enables the investigator to work with a manageable number. However, this grouping method has the disadvantage of obscuring individual scores. Thus all eighteen individual scores in the interval 20–22 might be 22 instead of 9 at 20, 5 at 21, and 4 at 22. We ignore this consideration when working with class intervals. Once the data are compressed by grouping, in all subsequent analysis and computation, we treat individual scores as if they were evenly distributed throughout the interval to which they belong.

Graphical representations of grouped frequency tables are very similar to those presented earlier for ungrouped data. One special kind of graph, the *frequency histogram,* is worthy of note. To construct a histogram from Table IV, as in Figure 3, the frequency of scores in each interval is represented by a rectangle with its center at the midpoint of the interval, its height equal to the frequency, and its width equal to the width of the interval's *limits*.

Here it is necessary to define and distinguish between two types of limits. When an interval is identified as 20–22, the limits reported are called *score limits*. They identify the lowest score and highest score that belong to the same interval. For purposes of mathematical treatment and graphical representation, it is common to use the *real limits* 19.5–22.5 to identify this same interval. The interval is represented as extending halfway to the scores immediately preceding and following. Such an interpretation is consistent with the way we usually report measurements.

For example, if we were measuring to the nearer meter, any measurement between 19.5 meters and 20.5 meters would be reported as 20 meters.

Calculating Averages

Tables and graphs can help us obtain considerable understanding of a set of scores. However, for many purposes, it is more desirable to try to represent the set of scores by a single number. When selecting a single number to represent a whole set of numbers, the first thing we usually think of is the average. As we have noted earlier, the scores we have been examining seem to cluster around a central point. It is this point of central tendency that statisticians identify when they report an average score. In statistics, there are several types of averages. Three are common in statistical analysis—the *mode,* the *median,* and the *mean*. Each is called a measure of central tendency. For the data with which we have been working in this article, the mode, median, and mean are close in value. This is not always the case. They may be appreciably different.

Mode.

The mode is quite easily identified from a frequency table or frequency graph. It is the score that occurs most frequently—in a sense, the most popular score. From the data presented in Table III, we can determine readily that the most frequent score is 16. Thus 16 is the mode of this set of scores. In the grouped data, the *crude mode* would be identified as the midpoint of the interval with the highest frequency. From either Table IV or Figure 3, we can see that the interval of highest frequency is 14–16. Hence 15 would be the crude mode of this distribution since it is the midpoint of the interval. Since individual scores have been obscured, we cannot be sure that it is actually the most frequent individual score, but it is our best estimate of the most frequent score.

Median.

The median is the middle score in a set of scores. The data in Table III contains 207 scores. The median is, then, the value of the 104th score. We find that the 104th score is among the 24 scores that all have a value of 15, and the median is therefore 15. If there had been an even number of scores, the median would have been reported as a figure that is halfway between the two middle scores if the two are different.

Since the set of data, presented in Table IV and Figure 3, has been condensed by grouping, we cannot work with individual scores and must find a new procedure for identifying the median. From Table IV, we find that the middle score, the 104th score, must fall in the interval 14–16. Since 79 scores out of the total number of 207 scores fall below this interval, we know that the median will be the 25th score in the interval (79 + 25 = 104). Since for grouped data we must assume even distribution within an interval, we will assume that the median lies $^{25}/_{64}$ of the width of the interval above its lowest boundary. We then multiply $^{26}/_{64}$ by 3 (the number of scores in the interval) for a product of $^{25}/_{64} \times 3 = 1.2$ (rounded to the nearest tenth). We now add 1.2 to the least-value endpoint of the interval 14–16. (We noted previously that this endpoint is 13.5.) Hence we have 13.5 + 1.2 = 15.7. The median, then, as calculated from the grouped data is 15.7.

Mean.

The mean is the most commonly used measure of central tendency, and it is the average most of us think of first. It is found by dividing the sum of all the individual scores by the number of scores in the set. The calculation of the sum of the scores can be shortened when a frequency table is available if we multiply each score by its frequency and then find the sum.

The mean for the data in Table III is calculated as follows:

Score		f		
3	×	1	=	3
4	×	2	=	8
5	×	2	=	10
6	×	3	=	18
7	×	6	=	42
8	×	9	=	72
9	×	7	=	63
10	×	7	=	70
11	×	12	=	132
12	×	20	=	240
13	×	10	=	130
14	×	12	=	168
15	×	24	=	360
16	×	28	=	448
17	×	22	=	374
18	×	13	=	234
19	×	11	=	209
20	×	9	=	180
21	×	5	=	105
22	×	4	=	88
		207		2,954

$2,954 \div 207 = 14.27 = 14.3$ = the mean

When computing the mean for grouped data, we assume even distribution of scores within each interval. We multiply the value of the midpoint of each interval by the frequency and divide the sum of the resulting number by the total number of scores.

The mean for the data in Table IV is calculated as follows:

Score		f		
3	×	3	=	9
6	×	11	=	66
9	×	23	=	207
12	×	42	=	504
15	×	64	=	960
18	×	46	=	828
21	×	18	=	378
		207		2,952

$2,952 \div 207 = 14.26 = 14.3$, the mean

Selecting Averages

The mode is used less frequently than either the median or the mean. It is useful only when we want to identify the number occurring most frequently in a set of numbers. As a matter of fact, if the mode is to be truly meaningful, one number in a set must occur quite a bit more frequently than any other number in the set. The advantage of the mode is that, like the median, it is easy to identify and understand. But the term "mode" is sometimes ambiguous, because there may be more than one score with the "highest frequency" (see Figure 4). Also, the mode is not reliable as an indication of central tendency, because the most popular score is not always near the center of a given distribution.

The chief advantage of the median is that it is not affected by extreme scores. The "average" income in a community, for example, is often more accurately reflected by the median than the mean, because the value of the median is not influenced by a few very high or very low incomes. The idea of the median is closely related to the concept of *percentiles*—a type of score students receive on certain standard tests in school. The median corresponds to the 50th percentile. Other percentiles can also be used in connection with tests. They are valuable as a basis for comparing individual scores with other scores in a distribution.

For most purposes, the mean is the best measure of central tendency. It is the only one of the three measures that depends upon the numerical value of each score in a distribution. It is a reliable indicator of central tendency because it always identifies the "balancing point" or "center of gravity" in the distribution. Since the mean lends itself better to mathematical computation, it is more suitable for deriving other statistical measures. For example, the means of two sets of data can be used to compute a mean for the combined set of data. This cannot be done with the mode or the median. However, the mean can give us information only about the central point in a distribution. To understand a set of scores more fully, we also need to know how the scores spread out around this central point. For this reason, statisticians develop measures of *dispersion* or *variability*.

The simplest measure of distribution is the *range,* which is defined as the difference between the highest and lowest scores in a distribution. The range of scores reported in Table III is easily identified as being 19. We simply subtract the lowest score (3) from the highest score (22). Since the range is sensitive only to the two extreme scores in a distribution, it is not considered a very satisfactory measure of dispersion. Its weakness is dramatized in the two distributions whose graphs are presented in Figures 4 and 5. In both distributions the mean is 12 and the range of scores is 12. Yet the distributions are obviously different.

As with the median and the mode, the weakness of the range is that it does not take into account the numerical value of each score. A natural measure of dispersion involving every score is the *average difference* between the individual scores and their mean. As we have seen, the mean serves as a "balancing point" for a distribution. Hence we can measure the deviations from the mean in terms of positive and negative differences. Deviations from the mean in one direction would be positive; in the other, negative.

The sum of the deviations from the mean is always zero. This is because the magnitude of negative differences always equals that of the positive differences. To avoid using negatives in the measurement of deviations, we use their *absolute values*. (The absolute value of a number disregards its positive or negative sign; the absolute value of any number x is represented as $|x|$.)

The average of the absolute values of the differences between individual scores and the mean is called the *mean deviation* and is a simple and accurate measure of dispersion. The calculation of the mean deviation for the distribution in Figure 4 is as follows:

f		Deviation from mean		
5	×	$\lvert 6-12 \rvert$	=	30
10	×	$\lvert 9-12 \rvert$	=	30
20	×	$\lvert 12-12 \rvert$	=	0
10	×	$\lvert 15-12 \rvert$	=	30
5	×	$\lvert 18-12 \rvert$	=	30
50				120

$120 \div 50 = 2.4$, the mean deviation

, the mean deviation.

A similar calculation for the distribution of test scores in Figure 5 would yield a mean deviation of 4.6. Therefore greater dispersion is indicated for the distribution in Figure 5.

The use of absolute values presents mathematical difficulties that can be avoided by using another measure. The positive and negative signs that led to the introduction of absolute values could also have been eliminated by squaring the deviations from the mean, since the square of a positive or negative number is always positive. Such a procedure preserves the descriptive

qualities of the mean deviation while providing a measure that is easier to handle mathematically. Hence statisticians prefer to use the *standard deviation,* indicated by the symbol "σ", as a measure of dispersion. The standard deviation is defined as the square root of the average squared deviation from the mean. Thus to calculate the standard deviation, we first find the average of the squares of the deviations. This number is called the *variance*. Suppose we wish to find the variance and standard deviation (σ) for the data in Table II.

Scores in Sixth Grade Class 6-1. The number of scores in this class is 35, and the mean is 15.7. Calculation of the variance and the standard deviation, σ, is as follows:

Standard Deviation of Test Scores for Class 6-1

Frequency	Deviation from Mean Squared			
2	×	$(12 - 15.7)^2$	=	27.38
3	×	$(13 - 15.7)^2$	=	21.87
4	×	$(14 - 15.7)^2$	=	11.56
6	×	$(15 - 15.7)^2$	=	2.94
9	×	$(16 - 15.7)^2$	=	0.81
5	×	$(17 - 15.7)^2$	=	8.45
3	×	$(18 - 15.7)^2$	=	15.87
2	×	$(19 - 15.7)^2$	=	21.78
1	×	$(20 - 15.7)^2$	=	18.49
35				129.15

129.15 ÷ 35 = 3.69, the variance $\sqrt{}$ (3.69) = 1.92, the standard deviation, σ

Similar calculations would yield a standard deviation of 4.1 for the data in Table IV, Scores in All Sixth-Grade Classes. If we compare these two measures of dispersion, we see that the scores in Table II the scores of Class 6-1, (the sample) are not so "spread out" as the scores in Table IV, the scores of all sixth grade classes (the population).

Interpreting the Data

Together, the mean and the standard deviation give us a reasonably clear picture of a distribution because they describe both its central tendency and its dispersion. Sometimes, if we know the general nature of the distribution, we need only these two numbers to reconstruct the distribution. For example, many sets of measurements have the *normal* distribution shown in Figure 6.

When a set of numbers "fits" such a standard distribution, we can determine approximately how many of the numbers fall within a given distance of the mean. For the normal distribution, 68.2 percent of the scores fall within one standard deviation of the mean. Thus, given the mean of 14.3 and the standard deviation of 4.1 for the set of scores in Table III, we could "predict" that 68.2 percent of the scores would be between 10.2 and 18.4. Since the distribution is actually given in Table III, we see that 152 of the 207 scores, or 68.1 percent, actually do fall in this interval. The

prediction is accurate because the number of scores is relatively large and they do fit the normal distribution.

The knowledge of such general models for distribution coupled with the laws of probability form the basis of *predictive statistics*. Both statistics and probability have to do with distributions, and it is upon this common focus that we capitalize in predictive statistics. In probability, the sample space (population) is known and we predict the composition of a set of outcomes (sample). In statistical inference, the sample (set of outcomes) is known, and we infer the composition of the population (sample space). Thus predictive statistics, also called statistical inference, can be thought of as an application of the laws of probability in reverse.

If we could be certain that the distribution of scores in a sample reflected exactly the distribution of scores in the population from which it was chosen, statistical inferences would be exact and simple to make. But even when a population is known, probability theory tells us that samples will not always be the same. The best we can hope for is that, if the sample is large enough and is carefully chosen, the sample characteristics will closely approximate those of the parent population.

Suppose we had only the data recorded in Table II for Class 6-1 and wished to estimate the mean score for all the English classes. By calculation, we have determined that the scores for Class 6-1 have a mean of 15.7 and a standard deviation of 1.92. Our best estimate of the mean for all classes would be equal to the mean of the sample, Class 6-1. But knowing that samples vary, we would hedge on this estimate. We would give a *confidence interval* within which we would expect the true mean of the population to fall. By assuming the total set of scores to be normally distributed and applying basic laws of probability, we could determine that there is a 95 percent chance that the population mean falls within 1.96 standard deviations or 3.69 score points (1.96 × 1.92). Thus there is a 95 percent chance that the population mean will be 15.7 plus or minus 1.8, or will be between 13.9 and 17.5. The 95 percent is a measure of the confidence or reliability we can place in our estimate. It means that 95 of every 100 populations from which a sample with the given characteristics might be chosen would have a mean within the determined interval. Because of the uncertainties involved in sampling, such an interval estimate, accompanied by a statement of the degree of confidence we can place in the estimate, is preferable to a single number approximation.

The process of establishing confidence intervals permits us to test hypotheses about a population. Modifications of this process permit us to make statistical comparisons of two samples drawn from the same population, to compare a sample to a known population, or to infer other characteristics of an unknown population.

F. Joe Crosswhite