

Chapter 1

Descriptive Statistics

1.1. Introduction

Statistics is a process for converting information into knowledge and making knowledge useful for the advancement of science. Many scientists use statistical methods to analyze their data in order to better understand a given research problem at hand and to help discover the unknown, and they regard statistical analysis to be an integral part of their research. Statistics as presented in this volume is a collection of analytic methods for scientific research and for practical applications.

The fundamental element of statistical analysis is the variable, the characteristic or outcome, which is measured or counted. In a human development study, for example, the variable of interest may be infant birthweight, length of gestation, birth order, sex of the child, or race of the mother. These, as all variables, by definition assume different values for different individuals. Birthweight and length of gestation are continuous variables in that they assume any of a continuum of values. Birth order is an ordinal variable because the values, the first birth, the second birth, etc., form a logical order. The race of an individual, which can assume any of several non-numeric values, is a nominal variable. A nominal variable with only two possible values is also called a dichotomous variable; sex is a dichotomous variable. Both nominal and ordinal variables are categorical variables

and are also called discrete variables because the values are distinct and do not fall on a continuum.

Variables associated with time, weight, or dimension are continuous variables. As a rule of thumb, a variable that can be *measured* as opposed to counted is a continuous variable. Due to the limitation of measuring scales, continuous variables are expressed in discrete units; weight is expressed in grams or pounds, length of gestation in months or weeks. The accuracy of the measurement of a continuous variable depends on the refinement of the scale: weight expressed in grams is more accurate than in kilograms. The following are additional examples of continuous variables:

- the time a clock stops;
- the distance of visibility on a specific day;
- the height of a college student;
- the level of cholesterol in a sample of blood.

Observed values of variables are expressed numerically so that the study data can be statistically analyzed. When a variable cannot be expressed numerically, such as a categorical variable, we record the number of observations falling into each category of the variable thus providing numbers for analysis, such as the number of students in a class, the number of dots on the faces of a pair of dice, the number of whales caught each year, etc.

Statistical data are a group of measurements, or observations, of some variable common to many people (or things) in a study, but they are different from personal data. They are non-personal. For example, age, sex, weight and height of a particular child are the child's personal data. However, these same measurements are detached from the child if the child is one of a group of children in a statistical study. These measurements become the data of one of many children and are treated in exactly the same way as those of any other child in the group. The fact that they relate to a particular child is disregarded.

Statistical data are different from personal data in yet other respects. First, they possess some properties not found in personal data. A group of observations, which forms a distribution, can be organized and summarized by what is known as descriptive statistics. Descriptive statistics describe the location, the dispersion and the pattern of the distribution by using numerals, tabulations, or graphics.

Further, the information derived from statistical analysis can be used to make inferences about some unknowns, such as the mean intelligence level of college students in California, the life expectancy of a newborn child,

etc. It is this property that makes statistics an important analytical tool in scientific research. This chapter is devoted to the discussion of descriptive statistics. In subsequent chapters, we shall present various methods of analysis for making statistical inference about unknowns.

In the study of statistical methods, we deal mostly with symbols and formulas, and little with numerical values. Numerical computations and numerical results are mainly for illustration, clarification, and introduction of concepts. While any symbol may be chosen to represent a variable, the process of representing variables in symbols and expressing procedures in formulas is fundamental to the study of statistical methods. When the symbols in a formula are understood, the formula can be applied to a large number of problems.

1.2. Measures of Location

Table 1.1 contains ages of 400 HIV positive men which will be used for illustration of descriptive statistics. These men constitute a sample; $n = 400$ is the sample size. The variable in this example is age, denoted by a symbol, Y . Thus y_1 denotes the age of the first man in the sample, y_2 the age of the second man, etc. As the identities of the 400 men are not of concern, the symbol y_1 may represent any one of the 400 ages, y_2 may represent any other age, etc. Here we let $y_1 = 49$ years, $y_2 = 42$ years, \dots , $y_{400} = 27$ years.

Statistical analysis can derive some meaning from a data set. The first question one might ask about the ages in Table 1.1 is: How old were these men on the average? Or, what value might represent the 400 ages? Here we are seeking a measure of the “central” value or “central tendency,” or the “location” of a group of observations. The following are the three most commonly used measures of location: The mean, the median, and the mode.

1.2.1. Mean

The mean, or the arithmetic mean (represented by \bar{y}), of a group of observations is the sum of the observations divided by the number of the observations. In formula, the mean of n observations, y_1, \dots, y_n , is given by

$$\bar{y} = \frac{y_1 + y_2 + \dots + y_n}{n} = \frac{1}{n} \left[\sum_{i=1}^n y_i \right]. \quad (1.1)$$

Table 1.1. Ages (last birthday in years) of 400 HIV positive men, San Francisco.

49	40	28	31	27	27	32	34	27	32
42	36	28	32	33	40	28	25	47	40
28	38	35	32	23	39	31	26	31	31
40	51	32	29	30	28	34	29	34	36
42	38	37	28	27	28	38	38	40	33
52	37	45	35	39	28	33	30	40	37
37	32	33	32	28	49	31	30	33	29
37	33	41	44	37	29	31	54	36	34
40	38	47	42	42	36	37	28	35	32
30	36	28	28	30	28	34	28	34	37
41	33	30	38	42	41	49	43	38	38
40	30	33	35	26	47	42	49	40	36
37	30	50	44	33	28	37	29	34	38
34	50	38	47	29	35	38	34	38	32
38	27	37	34	29	32	34	41	40	39
36	28	33	33	44	31	36	25	42	38
31	37	32	28	43	35	26	29	34	38
26	30	39	30	30	36	36	32	29	35
35	40	50	28	50	32	29	37	32	26
32	34	25	40	44	49	31	26	26	34
40	34	44	33	34	37	28	29	25	24
42	42	43	35	34	34	31	41	40	25
41	28	34	38	26	29	35	37	32	39
38	30	35	36	49	34	29	27	41	35
30	41	40	39	28	39	29	29	32	27
27	27	34	38	36	37	30	47	30	33
29	31	32	39	31	34	31	39	29	29
27	44	33	30	43	26	34	33	40	44
34	39	39	33	42	37	52	34	26	29
52	40	40	32	31	30	28	30	35	33
37	33	30	33	32	45	32	35	27	48
34	39	27	30	31	30	33	30	43	31
46	41	37	44	26	36	39	29	33	32
30	34	32	49	30	39	46	37	31	39
33	29	34	44	33	39	32	42	48	34
28	36	25	30	30	27	27	37	26	31
27	34	48	36	35	35	33	29	32	31
29	30	25	30	43	30	43	28	37	47
31	29	33	42	37	33	34	28	42	34
33	27	36	35	35	41	31	52	32	27

For the example data in Table 1.1:

$$\bar{y} = \frac{49 + 42 + \cdots + 27}{400} = \frac{13923}{400} = 34.81 \text{ years.}$$

Thus the mean age of the 400 men was 34.81 years.

1.2.1.1. *Some properties of mean*

(1) The mean also represents the balancing point of a group of observations, where the value of each observation is represented by a point on a horizontal scale. This balancing point is the center of gravity. Take the following set of four numbers:

$$(1, 4, 7, 8),$$

so that the mean is $\bar{y} = 5$. These numbers and the mean can be demonstrated graphically in Fig. 1.1.

If a unit weight is attached to each one of the four points (assuming the line is weightless), the line segment will be balanced at the point of the mean, $\bar{y} = 5$. This example illustrates an important property of the mean: the sum of the differences, called deviations, between each observation and the mean equals zero. In the present example,

$$(1 - 5) + (4 - 5) + (7 - 5) + (8 - 5) = 0.$$

An algebraic proof of this property is given in Sec. 1.7.

(2) If a constant a is added to each observation y_i , the mean is increased by the constant a . That is, the mean of $(a + y_1, a + y_2, \dots, a + y_n)$ is $a + \bar{y}$.

(3) If each observation y_i is multiplied by a constant b , the mean is multiplied by the constant b . That is, the mean of $(by_1, by_2, \dots, by_n)$ is $b\bar{y}$.

The last two properties can be combined to give a basic rule of the mean: The mean of a linear function of observations is equal to the linear function of the mean. That is, the mean of $(a + by_1, \dots, a + by_n)$ is $a + b\bar{y}$.

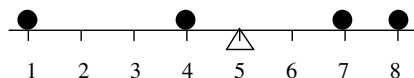


Fig. 1.1. Observed values and their mean.

1.2.2. Median

In the computation of the mean, the observations y_1, y_2, \dots, y_n need not be ordered. Formula (1.1) will yield the same value for the mean regardless of the value in the group each y_i represents. In the computation of the median, however, we require that the observations be ordered so that $y_1 < y_2 < \dots < y_n$. The median (M_d) of an ordered group of observations is the value that divides the observations into two equal subgroups. The same number of observations is above and below the median. If the number of observations n is odd,

$$M_d = \left(\frac{n+1}{2} \right) \text{th observation.} \quad (1.2)$$

To find the median of the five numbers (9, 4, 14, 10, 8), we first order them: (4, 8, 9, 10, 14) and then use formula (1.2) to locate the $(\frac{5+1}{2})$ th, or the third ordered value. That is $M_d = 9$.

When the number of observations n is even, the median is the average of two middle values. For example, the median of the $n = 4$ numbers (4, 8, 9, 14) is $\frac{(4+1)}{2} = 2.5$ th ordered value which is $\frac{(8+9)}{2} = 8.5$. Similarly, the median age of the 400 men in Table 1.1 is the average of the 200th and 201st ordered values, which is $M_d = 34.5$ years.

Remark 1.1. When n is even, any number between the two middle values has the property of the median. In the example (4, 8, 9, 14), 8.25 can be the median, as there are two numbers (9 and 14) greater than 8.25 and two numbers (4 and 8) less than 8.25. By convention, however, we use the average of the two middle values as the median.

Remark 1.2. The mean of a distribution is a more sensitive measure of location than the median is, in the sense that the mean is affected by every one of the observations, while the median may not be. In the example (4, 8, 9, 10, 14), both the mean and the median are equal to 9. If one of the observations is changed, the value of the mean changes as well. The mean of (4, 8, 9, 10, 14) is $\bar{y} = 9$ and the mean of (4, 8, 9, 10, 149) is $\bar{y} = 36$; but in both these cases the median remains $M_d = 9$. Which of the two measures is a better representation of the observations? The mean or the median? The choice is dependent on the nature of the data and the purpose of the study. If one needs a measure that takes into account every observed value in a sample, the mean is preferred. If a data set is likely to have a few extremely high values or a few extremely low values that will inappropriately influence the distribution, then the median is a better

choice. Most data dealing with personal income, for example, are described by the median because it is not unduly affected by the extremely wealthy or by the very poor and, therefore, it is more representative of the income of an entire group.

1.2.3. Mode

The mode (M_o) of a group of observations is the value of that observation which occurs most frequently. For example, the mode of the distribution (4, 8, 9, 9, 10, 14) is $M_o = 9$. The distribution in this example is also symmetric with respect to the value $y = 9$, as shown in the Fig. 1.2.

The mean, the median, and the mode all equal 9. Generally, in a symmetric (unimodal) distribution of observations, the mean, the median, and the mode coincide. However, the converse is not necessarily true. The three measures of location in a distribution may also coincide in an asymmetric distribution. In the group (4, 8, 9, 9, 11, 13), for example, the three measures of location all equal 9, but the distribution is asymmetric. Thus symmetry of a (unimodal) distribution implies that the three measures of location are equal, but equality of the three measures does not necessarily imply that the distribution is symmetric. See Fig. 1.3.

When a distribution is non-symmetric and the three measures of location are distinct, the distribution is skewed. A distribution may be skewed to the right (positively skewed) or skewed to the left (negatively skewed). A distribution skewed to the right usually has more extreme values to the right of the median and the mean is greater than the median. A distribution skewed to the left has the mean smaller than the median. The mode usually is on the other side of the median from the mean. For example,

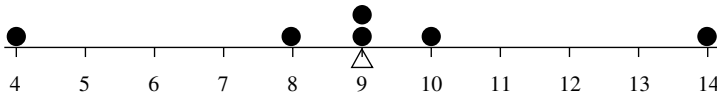


Fig. 1.2. The mean, the median, and the mode coincide in a symmetric distribution.

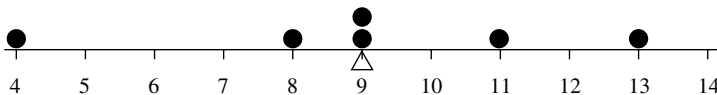


Fig. 1.3. The mean, the median, and the mode coincide in an asymmetric distribution.

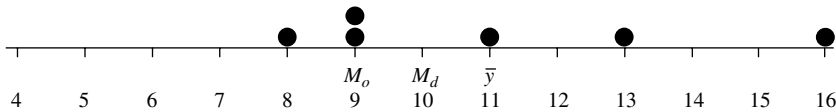


Fig. 1.4. The mean, the median, and the mode in a positively skewed distribution.

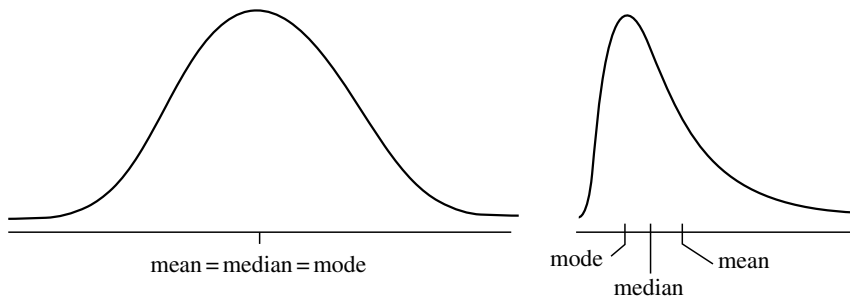


Fig. 1.5. A symmetric distribution and a positively skewed distribution.

the distribution of the numbers (8, 9, 9, 11, 13, 16) is skewed to the right, where the mean ($\bar{y} = 11$) is greater than the median ($M_d = 10$), and the median is greater than the mode ($M_o = 9$); the mean and the mode are on the two sides of the median, or $9 < 10 < 11$. See Fig. 1.4. Figure 1.5 shows a symmetric distribution and a skewed distribution (to the right) of continuous variables.

A distribution may have more than one mode. A distribution with one mode is called a unimodal distribution, while a distribution which has two or more modes is called a bimodal or multimodal distribution.

1.3. Measures of Dispersion

In the example (4, 8, 9, 9, 10, 14) in Fig. 1.2, the mean, the median, and the mode all equal 9, showing that the data are centered at the point $y = 9$. But measures of central tendency do not adequately describe a set of observations. For example, the mean of (7, 8, 9, 9, 10, 11) is also $\bar{y} = 9$, as are the median and the mode. But the two sets of data are quite different from one another. The numbers in (7, 8, 9, 9, 10, 11) are closer to one another, and to the mean, than are the numbers in (4, 8, 9, 9, 10, 14). The difference is in the dispersion or the variability of the data, which is not

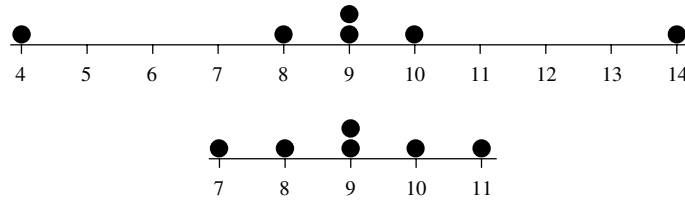


Fig. 1.6. Two distributions with the same central values but different dispersions.

reflected in the mean, median, or mode. Therefore, measures of dispersion are needed to describe a group of observations. See Fig. 1.6. There are several measures of dispersion, of which we discuss three.

1.3.1. Range

The range of a group of observations is the largest value minus the smallest value in the group, or

$$\text{Range} = \text{largest } y - \text{smallest } y. \quad (1.3)$$

A large value of the range indicates a large dispersion of the observations. In Fig. 1.6, the range of (4, 8, 9, 9, 10, 14) is $14 - 4 = 10$, while the range of (7, 8, 9, 9, 10, 11) is $11 - 7 = 4$. Therefore, the first set of observations has a larger dispersion than the second. The range is a very simple concept and is easy to determine, sometimes by inspection alone. But since it depends only on two extreme values and is not affected by any other values in a set of observations, the range is not a sensitive measure of dispersion. For example, the range of (4, 6, 9, 9, 12, 14) is also $14 - 4 = 10$. But the numbers in (4, 6, 9, 9, 12, 14) are more different from one another, and from the mean, than are the numbers in (4, 8, 9, 9, 10, 14). A more precise measure of dispersion of a set of data is the variance.

1.3.2. Variance and standard deviation

The variance of a sample of n observations is defined as the sum of squared deviations of observations from the mean divided by $n - 1$, or

$$S_Y^2 = \frac{1}{n - 1} \sum_{i=1}^n (y_i - \bar{y})^2. \quad (1.4)$$

The variance of a sample of observations takes into account the deviation of every y_i from the mean. The value of the variance is zero when all the n observations are identical; in this case there is no variability. The variance increases as the variability, or the deviations $y_i - \bar{y}$, increase. The variance of a sample of observations reflects quantitatively the degree of dispersion in the sample.

In formula (1.4), each derivation $(y_i - \bar{y})$ is squared to assure that the variance is a positive measure. If each deviation is not squared, the sum of the deviations $(y_i - \bar{y})$ is zero, as shown in Sec. 1.7 as a property of the mean. Also the fact that the sum of n squared deviations is divided by $n - 1$, not by n , may appear strange at first glance, but it is intuitively justifiable. There are only $n - 1$ *independent* deviations because any one of the deviations can be expressed in terms of the other $n - 1$ deviations. For example $(y_1 - \bar{y}) = -[(y_2 - \bar{y}) + \cdots + (y_n - \bar{y})]$. In three observations (1, 2, 6) with $\bar{y} = 3$ we have $(1 - 3) = -[(2 - 3) + (6 - 3)]$. More discussion of this point will be given in Sec. 3.4.4.

The standard deviation is defined as the (positive) square root of the variance, or

$$S_Y = \sqrt{\frac{1}{(n-1)} \left[\sum_{i=1}^n (y_i - \bar{y})^2 \right]}. \quad (1.5)$$

1.3.2.1. Some properties of variance and standard deviation

(1) If a constant a is added to each observation in the data set, the variance is not changed. In other words, a change of the location of a distribution has no effect on the dispersion of the distribution. In formula,

$$S_{a+Y}^2 = S_Y^2. \quad (1.6)$$

(2) If each observation is multiplied by a constant b , then the variance is multiplied by a factor b^2 , or

$$S_{bY}^2 = b^2 S_Y^2. \quad (1.7)$$

Both properties can be summarized as: the variance of a linear function of Y , $a + bY$, is equal to b^2 times the variance of Y , or

$$S_{a+bY}^2 = b^2 S_Y^2. \quad (1.8)$$

It follows that the standard deviation of a linear function of Y , $a + bY$, is equal to b times the standard deviation of Y .

Table 1.2.

Sample	R	S_y^2	S_y
(7, 8, 9, 9, 10, 11)	4	2.0	1.41
(4, 8, 9, 9, 10, 14)	10	10.4	3.22
(4, 6, 9, 9, 12, 14)	10	13.6	3.69

Table 1.2 shows the range, the variance, and the standard deviation of each of the three examples. The sample variance for the ages in Table 1.1 is

$$\begin{aligned} S_Y^2 &= \frac{1}{399} \left[\sum_{i=1}^{400} (y_i - 34.81)^2 \right] \\ &= \frac{1}{399} [(49 - 34.81)^2 + (42 - 34.81)^2 + \cdots + (27 - 34.81)^2] = 39.95, \end{aligned}$$

where y_i denotes the age and the standard deviation is $S_Y = \sqrt{39.95} = 6.24$ years.

1.3.2.2. Two alternative formulas for the variance

For easier computation, for combining two or more variances, or for the partition of sums of squares (discussed in later chapters), two alternative formulas of the variance are available:

$$S_Y^2 = \frac{1}{n-1} \left[\sum_{i=1}^n y_i^2 - n(\bar{y}^2) \right] \quad (1.9)$$

and

$$S_Y^2 = \frac{1}{n-1} \left[\sum_{i=1}^n y_i^2 - \frac{(\sum_{i=1}^n y_i)^2}{n} \right]. \quad (1.10)$$

1.3.3. Covariance $\text{Cov}(X, Y)$, or $S_{X,Y}$

The covariance between two variables, X and Y , is a measure of the variation of X and Y jointly. In a sample of n pairs of observations, $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, the formula for the covariance is

$$\text{Cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}). \quad (1.11)$$

The covariance may be positive, negative, or equal to zero. If Y increases as X increases, $\text{Cov}(X, Y) > 0$; if Y decreases as X increases, $\text{Cov}(X, Y) < 0$; if one of the variables is constant, or if the distribution of one variable remains the same for different values of the other variable, the covariance is zero. More discussion of the covariance will be given in Chap. 10 on correlation.

The computational formulas for the covariance are

$$\text{Cov}(X, Y) = \frac{1}{n-1} \left[\sum_i x_i y_i - \frac{1}{n} \sum x_i \sum y_i \right] = \frac{1}{n-1} \left(\sum x_i y_i - n\bar{x}\bar{y} \right). \quad (1.12)$$

1.4. Grouped Data

A data set generated from a research project usually is voluminous, unordered, confusing and even chaotic. Before attempting any statistical analysis, we need to rearrange and reduce the mass of data to a simple and compact form so that some meaningful information can be derived. The most effective form is the frequency distribution, which consists of a number of ordered intervals (or classes) and the number of observations falling into each interval, or the frequency. A data set in this form is called grouped data. A typical frequency distribution is shown in Table 1.3.

The following terms are associated with a frequency distribution:

Relative frequency $\left(\frac{f_i}{n}\right)$ is the proportion of observations in a specific interval.

Cumulative frequency (F_i) is the total number of observations up to the upper limit of a specific interval. That is, $F_i = f_1 + \dots + f_i$, for $i = 1, \dots, k$; and $F_k = n$.

Relative-cumulative frequency $\left(\frac{F_i}{n}\right)$ is the proportion of observations up to the upper limit of a specific interval.

Construction of a frequency distribution from a given set of data involves the determination of the number of intervals (k), the width of the interval (w), the limits of the intervals (L_i, L_{i+1}), and the corresponding frequencies (f_i). The number of intervals is usually set between 8 and 20 so that there will be enough intervals to show the general pattern of the distribution, but not too many so that the pattern is lost. The range of a data set divided by

Table 1.3. A frequency table.

Class or Interval	Midpoint	Frequency	Relative Frequency	Cumulative Frequency	Relative-cumulative Frequency
L_i to L_{i+1}	y_i	f_i	$\frac{f_i}{n}$	F_i	$\frac{F_i}{n}$
(1)	(2)	(3)	(4)	(5)	(6)
L_1 to L_2	y_1	f_1	$\frac{f_1}{n}$	F_1	$\frac{F_1}{n}$
L_2 to L_3	y_2	f_2	$\frac{f_2}{n}$	F_2	$\frac{F_2}{n}$
L_3 to L_4	y_3	f_3	$\frac{f_3}{n}$	F_3	$\frac{F_3}{n}$
.
.
.
L_k to L_{k+1}	y_k	f_k	$\frac{f_k}{n}$	F_k	$\frac{F_k}{n} = 1$
Total	—*	n	1.0	—*	—*

*The sum of this column has no meaning.

the number of intervals gives the approximate width of the interval. Generally, the width of the interval is an integer and is constant for all the k intervals. The first interval must include the smallest number in a data set and the last interval includes the largest. The midpoint of the interval should be a convenient number for computation. From the upper limit of the first interval, marking a length of w consecutively $k - 1$ times gives all the k intervals. Finally, tallying all the n numbers in a data set yields a frequency distribution. The entire grouping process can be done on a computer.

1.4.1. Computational formulas for grouped data

In computing the mean, the median, and the variance from a frequency distribution, we make the assumption that the observations in each interval are uniformly distributed in the interval and that the midpoint of the interval, denoted by y_i , represents all the observations in the interval. It follows that the formula of the mean is

$$\bar{y} = \frac{f_1 y_1 + f_2 y_2 + \cdots + f_k y_k}{\sum_{i=1}^k f_i} = \frac{1}{n} \left[\sum_{i=1}^k f_i y_i \right]. \quad (1.13)$$

The formula for the variance is

$$S_Y^2 = \frac{f_1(y_1 - \bar{y})^2 + f_2(y_2 - \bar{y})^2 + \cdots + f_k(y_k - \bar{y})^2}{(n - l)}$$

$$= \frac{1}{n - 1} \left[\sum_{i=1}^k f_i(y_i - \bar{y})^2 \right] \quad (1.14)$$

$$= \frac{1}{n - 1} \left[\sum_{i=1}^k f_i(y_i)^2 - \frac{(\sum_{i=1}^k f_i y_i)^2}{n} \right] \quad (1.15)$$

$$= \frac{1}{n - 1} \left[\sum_{i=1}^k f_i(y_i)^2 - n(\bar{y}^2) \right]. \quad (1.16)$$

The formula for the standard deviation for grouped data, as in the case for ungrouped data, is the (positive) square root of the variance, or

$$S_Y = \sqrt{S_Y^2}. \quad (1.17)$$

The formula of the covariance for the grouped data will be given in Chap. 9.

The formula for the median for a set of grouped data is

$$M_d = L + \left[\frac{\frac{n}{2} - F}{f} \right] w, \quad \text{where} \quad (1.18)$$

L = the lower limit of the interval containing the median value,

F = the cumulative frequency up to L ,

f = the frequency in the interval containing the median value, and

w = the width of the interval containing the median value.

The key to calculating the median from grouped data is the assumption of a uniform distribution of the f observations in the interval containing the median. By definition, the median is the value of the $(\frac{n}{2})$ th observation in a sample, or the value of the $(\frac{n}{2} - F)$ th observation in the interval containing the median. Under the uniform distribution assumption, the ratio of the difference $(M_d - L)$ to the width of the interval (w) is equal to the ratio of the two corresponding frequencies $(\frac{n}{2} - F)$ to f , that is,

$$\frac{M_d - L}{w} = \frac{\frac{n}{2} - F}{f}. \quad (1.19)$$

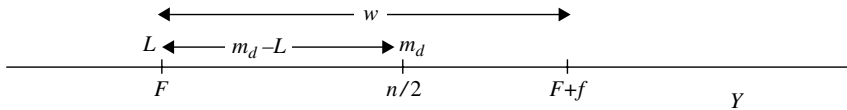


Fig. 1.7. Determination of the median for grouped data.

Table 1.4. Frequency distribution of ages of $n = 400$ men.

Interval (in years)	Midpoint	Frequency	Relative Frequency	Cumulative Frequency	Relative-cumulative Frequency
L_i to $L_i + 1$	y_i	f_i	$\frac{f_i}{n}$	F_i	$\frac{F_i}{n}$
(1)	(2)	(3)	(4)	(5)	(6)
22–26	24	10	0.0250	10	0.0250
26–30	28	78	0.1950	88	0.2200
30–34	32	103	0.2575	191	0.4775
34–38	36	88	0.2200	279	0.6975
38–42	40	62	0.1550	341	0.8525
42–46	44	31	0.0775	372	0.9300
46–50	48	18	0.0450	390	0.9750
50–54	52	9	0.0225	399	0.9975
54–58	56	1	0.0025	400	1.0000
Total	–*	400	1.0000	–*	–*

*This sum has no meaning.

Solving Eq. (1.19) for M_d yields the formula in (1.18). Figure 1.7 shows a graphic illustration of formula (1.19).

In deriving the formulas for the grouped data, we made an assumption of uniform distribution of f_i observations in each interval. This assumption generally is acceptable when the sample size is moderately large. For the special case of the 400 ages in Table 1.4, we have computed the mean, median, variance, and standard deviation from both grouped and ungrouped data, and presented them in Table 1.5. The two sets of numerical values differ only slightly from one another, indicating that the uniform distribution assumption is acceptable in this case.

1.5. Graphics

Graphics is another way of summarizing a mass of data in a simple form. A visual impression is direct and immediate and can convey much information

Table 1.5. Descriptive statistics of the ages of 400 men.

		Ungrouped (Table 1.1)	Grouped (Table 1.4)
Sample size	n	400	400
Mean	\bar{y}	35.31	35.3
Median	M_d	34.0	34.4
Variance	S_Y^2	39.95	40.39
Standard deviation	S_Y	6.24	6.4

quite effectively. Graphical representation has been extensively used to present summary information in many areas, including survey data, census figures, the national budget, and commercial advertising. The graphics most relevant to statistical analysis are the histogram, the frequency polygon and the cumulative frequency polygon.

1.5.1. Histogram

A histogram is a graphic representation of a (relative) frequency distribution. It consists of k rectangles, one for each interval, placed side by side on a horizontal axis. The base of each rectangle is the width of the interval, and the height is equal to the relative frequency ($\frac{f_i}{n}$). When the base of each rectangle is taken as a unit length, the area of each rectangle is $1 \times (\frac{f_i}{n})$, or equal to $\frac{f_i}{n}$, the relative frequency. It follows that the area of a histogram is the sum of the relative frequencies, or unity. If the height of each rectangle is equal to the frequency, f_i , then the area of a histogram is equal to the sample size, n . The two histograms are identical, except that the scales are different; the height of each rectangle is the relative frequency in one histogram, while it is the frequency f_i in the other. One may have two scales for a histogram — one on each side of the histogram. When relative frequencies are used, the name “relative frequency” histogram is more descriptive; the simple name “histogram” is used when there is no ambiguity. A relative frequency histogram has the advantage that it is consistent with the concept of probability (cf. Chap. 2), or proportion (cf. Fig. 1.8).

1.5.2. Relative frequency polygon

A relative frequency polygon also is a graphic representation of the distribution. It is constructed by line segments connecting midpoints at the top of neighboring rectangles in the relative frequency histogram. The polygon is

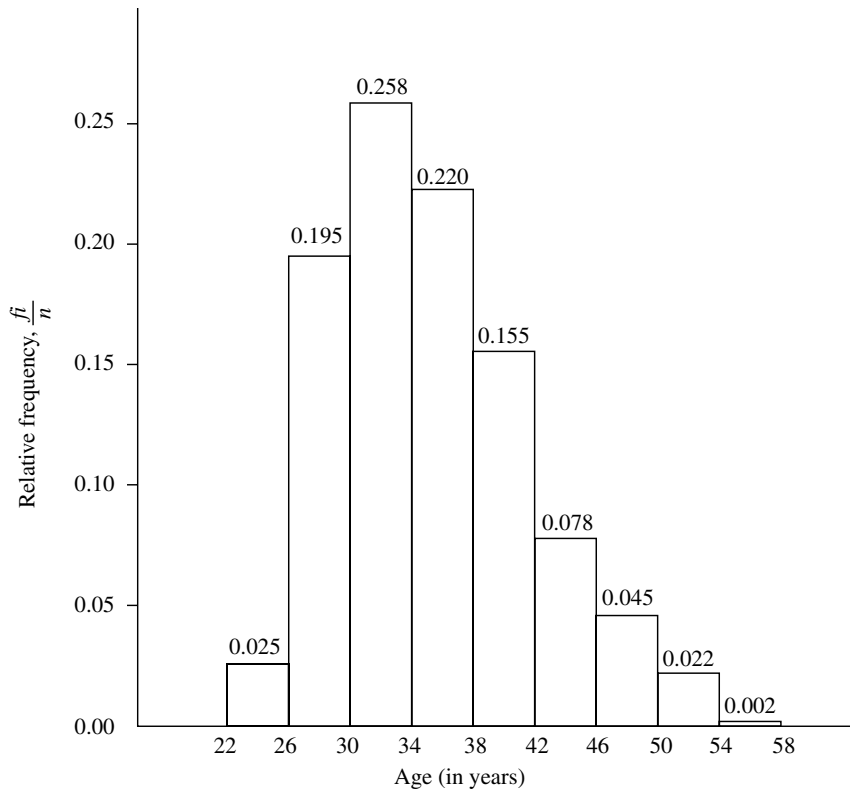


Fig. 1.8. Histogram: Age distribution of 400 HIV positive men.

completed by creating two additional intervals, each with a zero frequency, one below the first interval and the other above the last interval. The two midpoints on the horizontal axis are then connected by lines with the points in the first and the last intervals, respectively. These $k + 1$ lines plus the base line constitute a polygon as shown in Fig. 1.9. The area of a relative frequency polygon, as in the case of a histogram, is equal to unity.

1.5.3. Cumulative frequency polygon

A cumulative frequency polygon is yet another graphic representation of a frequency distribution or, more accurately, of a cumulative relative frequency distribution. It is constructed using lines connecting points over the upper limit of each interval with a height equal to the cumulative (relative)

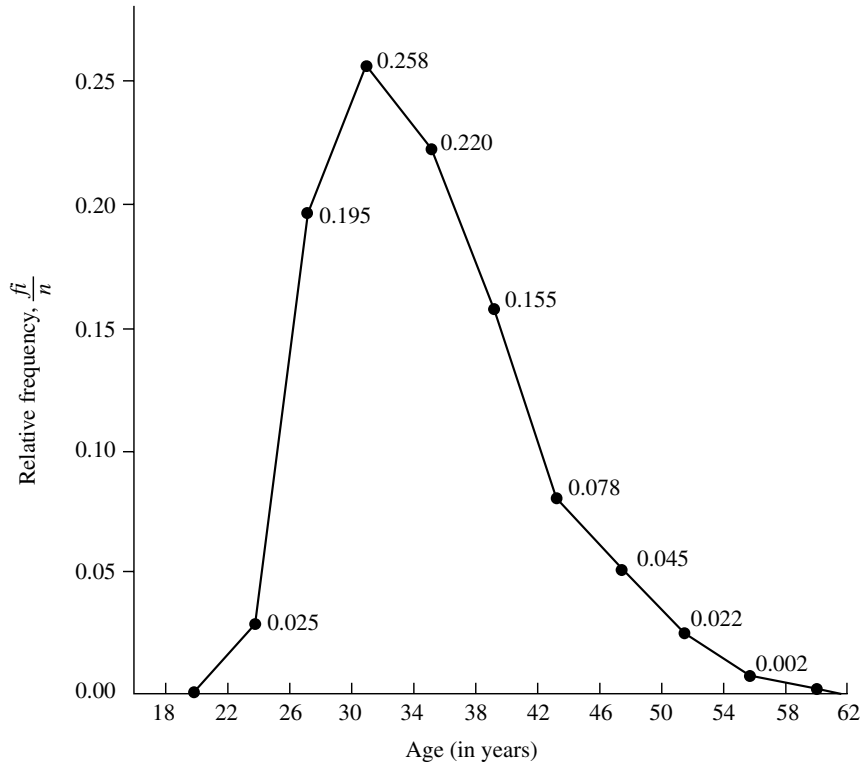


Fig. 1.9. Relative frequency polygon: Age distribution of 400 HIV positive men.

frequency of that interval ($\frac{F_i}{n}$). While for a histogram or a relative frequency polygon, the proportion of observations falling in a region is represented by an area, for a cumulative frequency polygon the proportion is represented by the height of a point on the polygon. The height of the point on the polygon over the upper limit of the last (the k th) interval is equal to unity, because it represents the proportion of all the observations in the sample. The numbers in Fig. 1.10 are the cumulative relative frequencies ($\frac{F_i}{n}$).

1.5.4. Percentiles and quartiles and interquartile range

The p th percentile in a sample is the point that divides the ordered observations into two pieces with p percent below and $(100 - p)$ percent above that point. When p is equal to 25, 50 or 75, the percentiles are called quartiles.

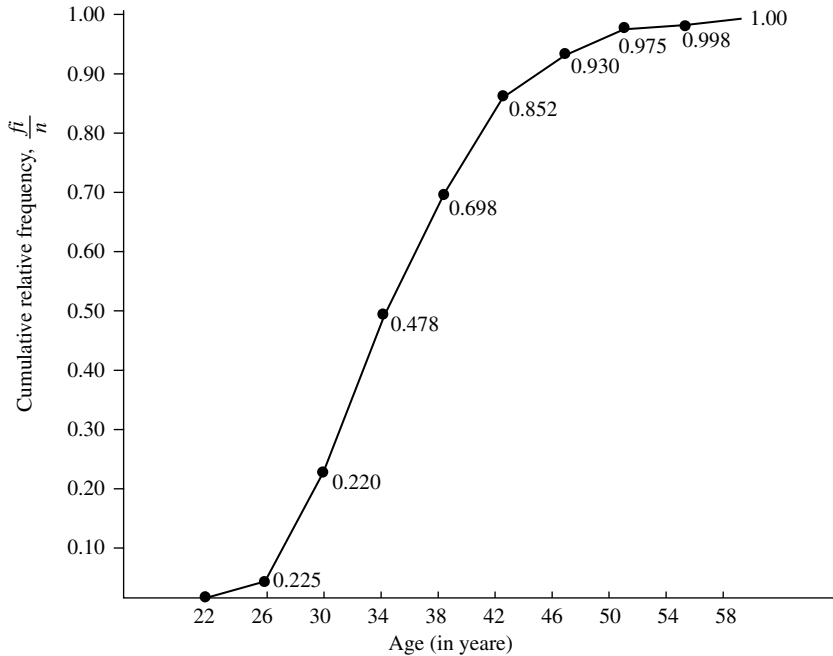


Fig. 1.10. Cumulative (relative) frequency polygon: Age distribution of 400 HIV positive men.

Thus the 25th percentile is the first quartile, the 50th percentile is the second quartile, and the 75th percentile is the third quartile. The 50th percentile, or the second quartile, is the median of a distribution. The distance between the first quartile and the third quartile is called the interquartile range, which is yet another measure of variability of observations in a sample. By definition, the interquartile range in a sample contains 50 percent of the observations. A large interquartile range indicates a high variability of the observations; a small interquartile range indicates a low variability.

Percentiles and quartiles usually are used in large samples and are determined from a frequency distribution. Their formulas are similar to the formula for the median. The first quartile, for example, is given by

$$Q_1 = L_1 + \left(\frac{\frac{n}{4} - F_1}{f_1} \right) \times w_1, \quad (1.20)$$

where L_1 denotes the lower limit of the interval where the first quartile is located, F_1 is the cumulative frequency up to L_1 , f_1 is the frequency in the

interval, and w_1 is the width of the interval. The formula for the third quartile is

$$Q_3 = L_3 + \left(\frac{\frac{3n}{4} - F_3}{f_3} \right) \times w_3 \quad (1.21)$$

where L_3, F_3, f_3 and w_3 are similarly defined as L_1, F_1, f_1 , and w_1 for the first quartile.

1.6. An Example

In Table 1.1 are ages (on last birthday) of a random sample of $n = 400$ men taken from census information of the Castro district in San Francisco. Each of these men was found to be antibody positive to the human immunodeficiency virus (HIV), the AIDS virus. We use this data set to illustrate descriptive statistics.

1. **Frequency distribution** The smallest and largest values in the group are 24 years and 54 years, respectively, giving the range $54 - 24 = 30$ years. Using $k = 9$ intervals and $w = 4$ years as the width of each interval, we let the first interval be (22–26) and mark from the point 26 a width of $w = 4$ years consecutively 8 times to determine all the 9 intervals as shown in Table 1.4, column (1). Tallying the $n = 400$ ages consecutively yields the frequencies in the 9 intervals and the frequency distribution in column (3). Note that the intervals are of exact age; two neighboring intervals overlap by one point. The first two intervals, for example, have a common point at (exactly) 26 years. Since age is a continuous variable, there are no persons whose ages are **exactly** 26 years at any given moment. Also, as the reported ages are those at last birthday, an age of 26 years should be entered in interval (26–30), an age of 30 years in interval (30–34), etc.

Relative frequencies, cumulative frequencies, and relative-cumulative frequencies are shown in columns (4)–(6) respectively.

2. **Mean, mode, and standard deviation** As the intervals in column (1) are exact ages, the midpoint in column (2) for each interval is equally distant from the lower and the upper limit of the interval. For example, the midpoint of the first interval is $24 = \frac{(22+26)}{2}$. Using formulas (1.13) and (1.15), we find the mean age $\bar{Y} = 35.3$ years, the variance $S_Y^2 = 40.39$, and hence the standard deviation $S_Y = 6.4$ years. The mode is $M_o = 32$ years.

3. Median To use formula (1.18) to compute the median, first look for the interval which contains the $(\frac{n}{2})$ th, or the 200th, ordered observation. The interval is (34–38). Next, identify the lower limit $L = 34$, the cumulative frequency $F = 191$, the frequency $f = 88$, and the width of the interval $w = 4$ years. Finally, substitute these values in formula (1.18) to obtain the median:

$$M_d = 34 + \left(\frac{200 - 191}{88} \right) \times 4 = 34.4 \text{ years.}$$

As the mode $M_o = 32$ years, the median $M_d = 34.4$ years and the mean $\bar{Y} = 35.3$ years, and $M_o < M_d < \bar{Y}$, the distribution is skewed to the right, or positively skewed.

4. Quartiles and interquartile range From the cumulative frequencies in column (5), we find that the first quartile Q_1 is located in the interval (30–34), and the third quartile Q_3 is in the interval (38–42). Using formula (1.20), we compute the first quartile

$$Q_1 = 30 + \left(\frac{100 - 88}{103} \right) \times 4 = 30.5 \text{ years.}$$

Using formula (1.21), we have the third quartile:

$$Q_3 = 38 + \left(\frac{300 - 279}{62} \right) \times 4 = 39.4 \text{ years.}$$

Therefore, the interquartile range is from 30.5 years to 39.4 years, which contains 50 percent of the ages in the sample.

5. Histogram, frequency polygon, and cumulative frequency polygon Following the instructions in Sec. 1.5, a histogram, a relative frequency polygon and a cumulative relative polygon have been prepared (Figs. 1.8–1.10, respectively).

The histogram in Fig. 1.8 and the relative frequency polygon in Fig. 1.9 are subject to similar interpretation. First of all, the height of each rectangle in the histogram and the height of each point in the polygon are equal, and equal to the relative frequency in the interval, denoted by a number. The area of the histogram and the area under the polygon are both equal to unity. Further, the area to the left (or right) of a point is the proportion of men (observations) who were younger (or older) than the age represented by the point. And finally, the area between any two points is the proportion of men (observations) whose ages were between the ages represented by the two points. For example, the area to the left of 30 is 0.22 ($= 0.025 + 0.195$),

which means that 22 percent of the men in the sample were younger than 30 years of (exact) age. Similarly, 2.4 percent ($= 0.022 + 0.002$) of the men were older than 50 years of age. Finally, 63.3 percent ($= 0.258 + 0.220 + 0.155$) of them were between 30 and 42 years of age.

In the cumulative (relative) frequency polygon in Fig. 1.10, the height of each point, denoted by a number, is the cumulative relative frequency up to the upper limit of the corresponding interval. Generally, the height of any point on a line segment on the polygon represents the proportions of men (observations) who were younger than the age represented by the point on the horizontal axis.

As the sample size n and the number of intervals k increase indefinitely, the proportions become probabilities, each rectangle in the histogram reduces to a (vertical) straight line (which is known as the density function), the relative frequency polygon approaches a smooth curve, and the cumulative frequency polygon tends to an S -shaped curve (known as the distribution function). The probability density, the smooth curve, and the S -shaped curve, each describes the probability distribution of a continuous variable. Thus, the graphics, as presented in this chapter, are an important first step to the understanding of the probability distribution.

6. A final remark This chapter contains only a few of the many possible techniques included in descriptive statistics. The use of descriptive statistics is more of an art than a science. The choice of the specific descriptive method often depends on the subject matter under study, the focus of the investigator, or the impression one wishes to convey. The choice is semantic; the availability of a large variety of descriptive statistics allows a wide range of ways to express the same information. British statesman Benjamin Disraeli once remarked — “there are lies, damned lies and statistics.” Quite possibly, the Victorian Prime Minister was puzzled by the number of ways one can present information with statistics. Today, it may be equally appropriate to say “there are lies, damned lies and politics.”

1.7. Proofs of the Results in this Chapter

In the following proofs \sum is written for $\sum_{i=1}^n$ for simplicity.

1. The sum of the deviations of each observation from the mean equals zero.

$$\sum (y_i - \bar{y}) = \sum y_i - \sum \bar{y} = \sum y_i - n\bar{y} = \sum y_i - \sum y_i = 0. \quad (1.22)$$

2. The mean of a linear function of a variable is equal to the linear function of the variable. Or, the mean of $a+by$ is $a + b\bar{y}$.

$$\frac{1}{n} \sum (a + by_i) = \frac{1}{n} \left(\sum a + b \sum y_i \right) = \frac{1}{n} (na + bn\bar{y}) = a + b\bar{y}. \quad (1.23)$$

Note: $\sum a = a + a + \cdots + a = na$.

3. The variance of a linear function of a variable, $a + bY$, is equal to b^2 times the variance of the variable Y .

$$\begin{aligned} S_{a+by}^2 &= \frac{1}{n-1} \sum [(a + by_i) - (a + b\bar{y})]^2 = \frac{1}{n-1} \sum [(by_i - b\bar{y})]^2 \\ &= \frac{b^2}{n-1} \sum [(y_i - \bar{y})^2] = b^2 S_Y^2. \end{aligned} \quad (1.24)$$

4. $\sum (y_i - \bar{y})^2 = \sum y_i^2 - \frac{(\sum y_i)^2}{n}$. Since

$$(y_i - \bar{y})^2 = y_i^2 - 2\bar{y}y_i + \bar{y}^2, \quad (1.25)$$

$$\begin{aligned} \sum (y_i - \bar{y})^2 &= \sum y_i^2 - 2\bar{y} \sum y_i + n\bar{y}^2 \\ &= \sum y_i^2 - 2n\bar{y}^2 + n\bar{y}^2 \\ &= \sum y_i^2 - n\bar{y}^2 \quad (\text{cf. formula (1.9)}) \\ &= \sum y_i^2 - n \left(\frac{\sum y_i}{n} \right)^2 \\ &= \sum y_i^2 - \frac{(\sum y_i)^2}{n}. \quad (\text{cf. formula (1.10)}) \end{aligned}$$

1.8. Exercises and Problems

1. Prepare a frequency distribution of the data in Table 1.1, using $w = 4$ with the first interval (20–24). Do all the work as demonstrated in Sec. 1.6.
2. Take a sample of $n = 9$ ages (with replacement) from Table 1.1 and find the mean, the median and the variance. [Be sure to adjust for the fact that Table 1.1 includes age at last birthday.]
3. Take a sample of $n = 9$ ages (with replacement) from Table 1.1, compute the mean and denote it by \bar{y}_1 . Take a second sample of $n = 9$ ages from Table 1.1 and denote the mean by \bar{y}_2 . Take $N = 101$ such samples and

compute the sample means $(\bar{y}_1, \bar{y}_2, \dots, \bar{y}_{101})$. [Be sure to express the means in exact age.]

4. **Continuation.** Compute the mean \bar{y} and the variance $S_{\bar{y}}^2$ of the $N = 101$ sample means in Problem 3. Compare the mean \bar{y} with the mean age $\bar{Y} = 35.3$ years and the product $nS_{\bar{y}}^2$ with the variance $S_Y^2 = 40.39$ of the 400 ages.