

## Chapter 1

# Descriptive Statistics

The statistical analyses in practice usually include two parts: statistical description and statistical inference. Statistical description is a kind of fundamental work for statistical inference, which describes the feature of the sample. The main forms for description are tables (such as frequency table), plots (such as scatter diagram, histogram) and numerical indexes (such as mean, standard deviation).

### 1.1 Variables and Data

#### 1.1.1 *Types of variables*

Variables are used to describe the properties of individuals in statistics. Different types of variables have different types of distributions and hence the statistical methods being used might be different. It is important to identify the types of variables before dealing with the data.

##### 1.1.1.1 *Continuous variable*

They are the variables whose values can be obtained through measurement such as the height, weight, blood pressure, pulse and blood count of the individuals. Limited by the precision of measurement, the variables such as height and weight can take some values of real number but not all indeed, and the variables such as pulse and blood count can take values of integral number only. However, for convenience in theoretical study, they are regarded as continuous variables taking values in a continuous interval on the axis of real number. Sometimes, the observed values of such kind of variables are called measurement data.

### 1.1.1.2 Discrete variable

Some properties can only be described qualitatively with several mutually excluded categories, such as gender, occupation and effect of medicine (positive or negative). The variable for gender can only take a “value” either “male” or “female”; the variable of occupation may take a “value” among several categories (worker, farmer, sales man and soldier etc.). This kind of variables is called categorical variables or nominal variables.

**Example 1.1** The variable for gender can be defined with a binary variable  $X$ .

$$X = \begin{cases} 0 & \text{Female,} \\ 1 & \text{Male.} \end{cases}$$

In general, the variables taking values in a set of countable numbers are called discrete variables. Binary variable is a simplest special case of it.

The number of individuals within certain category is often counted, and it is called frequency so that the data of discrete variable is sometimes called count data.

**Example 1.2** In the sample of 108 patients, there are 63 males and 45 females. If a binary variable  $X$  is defined for gender as Example 1.1, the sum of  $X$  for the 108 patients is the number of males (63).

In general, the frequency of certain category is equivalent to the sum of a binary variable.

### 1.1.1.3 Ordinal variable

Some measurement can only result in a semi-quantitative outcome. For instance,  $-$ ,  $\pm$ ,  $+$ ,  $++$ ,  $+++$  are quite often used to indicate different ranks in clinic. For some properties, there naturally exist ranks among different categories. For instance, cure, effective, ineffective and worse are used to describe the level of drug effect. An ordinal variable can be defined for this kind of properties taking values among 1, 2, 3, ... for rank, but not for the exact quantitative measurement.

The frequencies of ordinal variable is sometimes called ranked data.

### 1.1.2 Structure and feature of data

Any outcome of experiment or observation should be expressed with numerical data for statistical analysis. Most outcomes in medical research could be expressed through a data structure similar to Table 1.1, where 7 recorded items of 100 patients are given by a matrix with 100 rows and 7 columns. This is a basic format for data input in most of statistical software such as SAS, SPSS and BMDP.

#### 1.1.2.1 Basic observed unit

It is the basic unit for data collection determined by the purpose of research. For instance, if the systolic pressure and diastolic pressure are measured at a fixed time point after treatment then a patient is defined as an observed unit; otherwise, if the systolic pressure and diastolic pressure are measured at 3 time points after treatment (say, week 1, week 2 and week 4), then each patient is regarded as 3 observed units since the condition of each patient changes with time.

#### 1.1.2.2 Recording item

The recording items used for statistical analysis usually consist of 3 parts: group, response variables and covariates. Columns 2–8 of Table 1.1 show a  $100 \times 7$  matrix corresponding to 7 recording items, of which treatment is a variable for grouping, systolic pressure, diastolic pressure, ECG and effectiveness are response variables, and age and gender are covariates.

Table 1.1 The post-treatment clinical records of 100 hypertension patients.

No. (1)	Age (years) (2)	Gender (3)	Treatment (4)	Systolic pressure (kPa) (5)	Diastolic pressure (kPa) (6)	ECG (7)	Effectiveness (8)
1	37	Male	Drug A	18.67	11.47	Normal	Prominent
2	45	Female	Control	20.00	12.53	Normal	Effect
3	43	Male	Drug B	17.33	10.93	Normal	Effect
4	59	Female	Control	22.67	14.67	Abnormal	No effect
...	...	...	...	...	...	...	...
100	54	Female	Drug B	16.80	11.73	Normal	Effect

## 1.2 Frequency Table and Histogram

Frequency table and histogram are not only fairly useful for description of sample data but also the intuitive basement of the important concept of probability distribution.

### 1.2.1 Frequency table

As mentioned before, in a set of sample, the number of appearing times for certain event is frequency. For a complete list of mutually exclusive events, the table putting the corresponding frequencies together is called a frequency table.

#### 1.2.1.1 Discrete-type frequency table

To a discrete variable, the completely and mutually exclusive events are just the possible values or categories of that variable. Based on the data of Example 1.3, two frequency tables are given in Tables 1.2 and 1.3, where the ratio between the frequency and the total number is called relative frequency (for simple, which is also called frequency when it is not confused). The sum of all relative frequencies must be 100% (in practice, sometimes not being exactly 100% due to rounding error). The cumulative frequencies and cumulative relative frequencies are the results by successively cumulating the frequencies and relative frequencies respectively.

It is similar for ordinal variables. For instance, Table 1.4 is a frequency table for the results of certain semi-quantitative test among 150 patients; Table 1.5 is a frequency table for the treatment effect after their taking certain medicine.

Table 1.2 The frequency table for gender of 108 patients.

Categories of gender	Frequency	Relative frequency (%)	Cumulative frequency	Cumulative relative frequency (%)
Female	45	41.7	45	41.7
Male	63	58.3	108	100.0
Total	108	100.0		

Table 1.3 The frequency table for occupation of 108 patients.

Categories of occupation	Frequency	Relative frequency (%)	Cumulative frequency	Cumulative relative frequency (%)
Labour	28	25.9	28	25.9
Farmer	23	21.3	51	47.2
Business	24	22.2	75	69.4
Student	18	16.7	93	86.1
Soldier	15	13.9	108	100.0
Total	108	100.0		

Table 1.4 The frequency table for the results of certain semi-quantitative test among 150 patients.

Categories of the results	Frequency	Relative frequency (%)	Cumulative frequency	Cumulative relative frequency (%)
–	80	53.3	80	53.3
±	20	13.3	100	66.6
+	25	16.7	125	83.3
++	15	10.0	140	93.3
+++	10	6.7	150	100.0
Total	150	100.0		

Table 1.5 The frequency table for the treatment effect of certain medicine.

Categories of effectiveness	Frequency	Relative frequency (%)	Cumulative frequency	Cumulative relative frequency (%)
Cure	65	43.3	65	43.3
Effect	45	30.3	110	73.6
No effect	25	16.7	135	90.0
Worse	15	10.0	150	100.0
Total	150	100.0		

### 1.2.1.2 Continuous type frequency table

To continuous variable, the general method for establishing a frequency table could be learnt from the following example.

**Example 1.3** 120 normal male adults were randomly selected from the residents of a county. Their red cell counts ( $10^{12}/L$ ) were observed and listed as the follows:

5.12	5.13	4.58	4.31	4.09	4.41	4.33	4.58	4.24	5.45	4.32	4.84
4.91	5.14	5.25	4.89	4.79	4.90	5.09	4.04	5.14	5.46	4.66	4.20
4.21	3.73	5.17	5.79	5.46	4.49	4.85	5.28	4.78	4.32	4.94	5.21
4.68	5.09	4.68	4.91	5.13	5.26	3.84	4.17	4.56	3.52	6.00	4.05
4.92	4.87	4.28	4.46	5.03	5.69	5.25	4.56	5.53	4.58	4.86	4.97
4.70	4.28	4.37	5.33	4.78	4.75	5.39	5.27	4.89	6.18	4.13	5.22
4.44	4.13	4.43	4.02	5.86	5.12	5.36	3.86	4.68	5.48	5.31	4.53
4.83	4.11	3.29	4.18	4.13	4.06	3.42	4.68	4.52	5.19	3.70	5.51
4.64	4.92	4.93	4.90	3.92	5.04	4.70	4.54	3.95	4.40	4.31	3.77
4.16	4.58	5.35	3.71	5.27	4.52	5.21	4.37	4.80	4.75	3.86	5.69

Please try to establish a frequency table for this set of data.

#### Solution

##### (1) Range $R$

The difference between the maximum and minimum of the data set is called range. In our example, maximum = 6.18, minimum = 3.29, the range is  $R = 6.18 - 3.29 = 2.89$ .

##### (2) Length of sub-intervals $i$

Divide the whole range into 8–15 sub-intervals. For convenience, take one tenth of the range firstly, and then slightly adjust to a easy number. In our example,  $R/10 = 2.89/10 = 0.289 \approx 0.30$ , then let  $i = 0.30$ .

##### (3) Work out the list of sub-intervals

First of all, take a number slightly less than the minimum as the lower limit of the first sub-interval, say 3.20, such that its upper limit is  $3.20 + 0.30 = 3.50$ ; take 3.50 as the lower limit of the second sub-interval such that its upper limit is  $3.50 + 0.30 = 3.80$ ; ... Due to that the upper limit of the former sub-interval is equal to the lower limit of the later one, for convenience, the upper limits are open and not shown except the last sub-interval, hence the list of sub-intervals are 3.20~, 3.50~, 3.80~, ..., 5.60~ and 5.90~6.20 (column 1 of Table 1.6).

Table 1.6 The frequency table based on the data set of red cell counts of 120 normal male adults.

Sub-interval (1)	Mark	Frequency (2)	Relative frequency (%) (3)	Cumulative frequency (4)	Cumulative relative frequency (%) (5)
3.20~	丁	2	1.7	2	1.7
3.50~	正	5	4.2	7	5.9
3.80~	正正	10	8.3	17	14.2
4.10~	正正正正	19	15.8	36	30.0
4.40~	正正正正丁	23	19.2	59	49.2
4.70~	正正正正正	24	20.0	83	69.2
5.00~	正正正正正一	21	17.5	104	83.7
5.30~	正正一一	11	9.2	115	95.9
5.60~	正	4	3.3	119	99.2
5.90~6.20	一	1	1.7	120	100.0
Total		120	100.0		

**(4) Read, mark and count to get frequencies**

Read over the data and write the five strokes of the Chinese character “正” one by one to mark and count the number of individuals corresponding to each sub-intervals (column 2 of Table 1.6).

(5) Calculate the frequencies, relative frequencies and cumulative ones (columns 3–5 of Table 1.6).

**1.2.2 Frequency plot and histogram**

To present the frequency table intuitively, a frequency plot within a coordinate system can be used, where the horizontal axe refers to “various situations” of the variable and the vertical axe refers to the corresponding frequencies or frequency densities.

**1.2.2.1 Frequency plot for discrete variable — bar chart**

To a discrete variable, one can use the points on the horizontal axe to express different categories or their related values; and plot vertical line segments on these points, of which the lengths express the frequencies or relative frequencies of the corresponding categories (Figs. 1.1 and 1.2). Such kind of frequency plot is called bar chart.

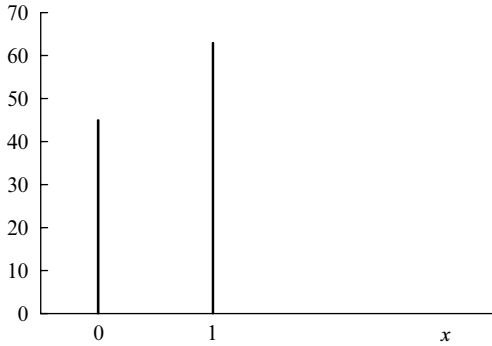


Fig. 1.1 The frequency plot for gender of 108 patients.  $x$ : gender, 0: female, 1: male.

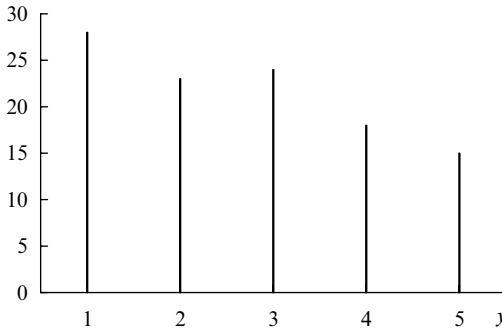


Fig. 1.2 The frequency plot for occupation of 108 patients.  $y$ : occupation, 1: labour, 2: farmer, 3: business, 4: student, 5: soldier.

### 1.2.2.2 Frequency plot for continuous variable — histogram

To a continuous variable, one can use the sub-intervals with equal length on the horizontal axis to express different situations of the variable; and plot vertical rectangles on these intervals, of which the heights express the frequencies related to the sub-intervals (Fig. 1.3(a)). However, when the lengths of the sub-intervals are not equal (for instance, the age intervals  $0\sim, 1\sim, 5\sim, 10\sim, 15\sim, \dots$ ), the heights cannot be used to express the frequencies.

Alternatively, one would use the areas of the rectangles to express the relative frequencies, and call such kind of plot histogram. The height of any rectangle in a histogram is neither the frequency nor the relative

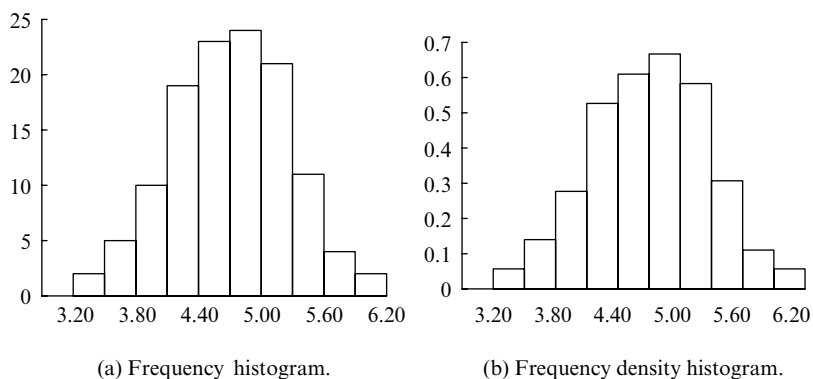


Fig. 1.3 Histograms plotted on the basis of the frequency table for the data set of red cell counts ( $10^{12}/L$ ) of 120 normal male adults.

frequency, but the ratio of the relative frequency to the length of the sub-interval. Such kind of histogram is called frequency density histogram, of which the total area of all the rectangles is equal to 1 or 100%. The frequency density histogram can be used no matter the lengths of the sub-intervals are equal or not.

Both of the frequency histogram and the frequency density histogram reflect the chances of various values taken by a continuous variable. The histograms in Fig. 1.3 appear to be symmetric, higher around center and shorter on two sides, which indicate that the red cell counts of normal male adults may be higher or lower with about equal chances, but mostly around the median level. Many histograms in practical problems look like this type. However, there are some other types as well. For instance, the frequency histogram of hair mercury for the residents of a city is given in Fig. 1.4; the frequency histogram of age for a group of male patients with lung cancer is given in Fig. 1.5; and the frequency histogram of the scores suggested by a group of patients for the importance of a specific item in evaluation of quality of life is given in Fig. 1.6. One can see, Figs. 1.4 and 1.5 are higher around center and shorter on two sides but not symmetric, of which the shape is usually called skew. The tail on the positive side appears longer in Fig 1.4 and hence it is called positive skew, and the tail on the negative side appears longer in Fig. 1.5 and hence it is called negative skew. The histogram in Fig. 1.6 appears shorter around center and higher on two sides, of which

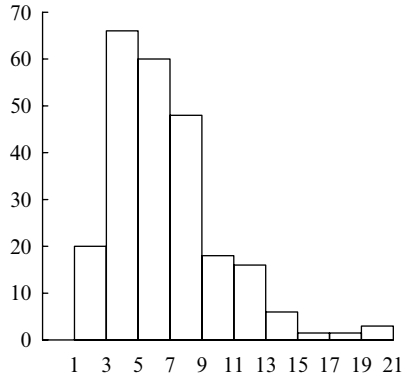


Fig. 1.4 Frequency histogram of hair mercury for the residents of a city.

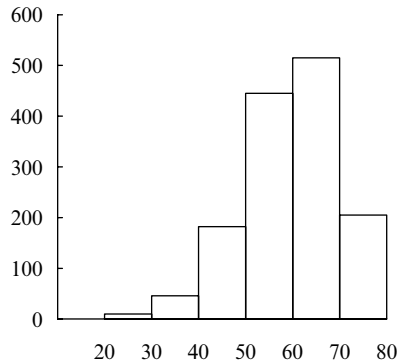


Fig. 1.5 Frequency histogram of age for a group of male lung cancer patients.

the shape looks like a hook. Various shapes of the histograms are important for us to learn the distributions of continuous variables.

### 1.2.2.3 Frequency plot for ordinal variable — bar chart

The distances between successive ranks of an ordinal variable are usually un-equal or unknown so that a bar chart instead of a histogram is used for frequency plot. For instance, the effect of a treatment can be described with 4 ranks, cure, effect, no effect and worse, and the corresponding frequencies can be expressed with 4 bars standing up the horizontal axis as a bar chart for discrete variable.

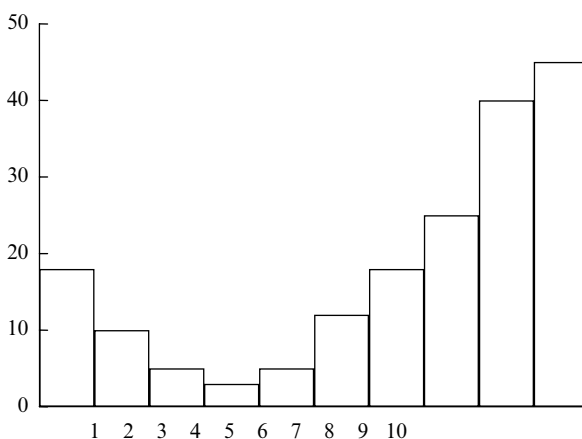


Fig. 1.6 Frequency histogram of the scores suggested by a group of patients for the importance of a specific item in evaluation of quality of life.

### 1.3 Measurement for Average Level of a Sample

In addition to frequency table and histogram, numerical characteristics are also used for statistical description. To continuous variables, two often-used characteristics are average level and variation. Depending on the type of distribution, different measurements could be used to describe the average level of a group of varied values.

#### 1.3.1 Arithmetic mean

When the histogram looks symmetric, the one can well-represent the average level is the arithmetic mean, or mean or average for brief, which is equal to the quotient of dividing the sum of observed values by the total number of individuals.

##### 1.3.1.1 Raw data based approach

Denote the observed values of the individuals with  $x_1, x_2, \dots, x_n$  and the arithmetic mean with  $\bar{x}$ , then

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \sum_{i=1}^n \frac{x_i}{n}. \quad (1.1)$$

Whenever it is not confused,  $\sum_{i=1}^n x_i$  could be simplified as  $\sum x_i$  or even  $\sum x$ . (1.1) is an approach to calculate the mean directly on the basis of the raw data.

1.3.1.2 Frequency table based approach

When the raw data are not available, the frequency table can be used to calculate the mean approximately. Usually the mid-values of the sub-intervals are taken as the representative values. If one wants to calculate the mean based on Table 1.6, then follows Table 1.7, the mean is

$$\begin{aligned} \bar{x} &= 3.35 \times 0.017 + 3.65 \times 0.042 + \dots + 6.05 \times 0.017 \\ &= 0.0569 + 0.1533 + \dots + 0.1028 = 4.7057. \end{aligned}$$

Obviously, it is approximate to the mean obtained on the basis of raw data where  $\bar{x} = 4.7167$ .

The formula for the above approach can be expressed as

$$\bar{x} = \sum_{i=1}^n \left( \frac{f_i}{n} \right) x_i \tag{1.2}$$

where  $f_i$  and  $x_i$  are the relative frequency and mid-value of the  $i$ th sub-interval,  $n$  is the total sample size. One can see from the process of above

Table 1.7 The operation of weighted average based on a frequency table.

Sub-interval (1)	Mid-value ( $x$ ) (2)	Frequency ( $f$ ) (3)	Relative frequency ( $f/n$ ) (4) = (3)/120	Mid-value $\times$ Relative frequency (5) = (2) $\times$ (4)
3.20~	3.35	2	0.017	0.0569
3.50~	3.65	5	0.042	0.1533
3.80~	3.95	10	0.083	0.3278
4.10~	4.25	19	0.158	0.6715
4.40~	4.55	23	0.192	0.8327
4.70~	4.85	24	0.200	0.9700
5.00~	5.15	21	0.175	0.9013
5.30~	5.45	11	0.092	0.9537
5.60~	5.75	4	0.033	0.1897
5.90~6.20	6.05	1	0.017	0.1028
Total		120	1	4.7057

calculation, the mid-value  $x_6 = 4.85$  is multiplied by a bigger frequency  $f_6/n = 20.0\%$  hence the contribution of  $x_6$  is bigger. Such a way that the mid-values are not equally dealt with in the process of making average is called weighted average, and the result is called weighted mean. The element in (1.2) reflecting the importance of the mid-value  $x_i$  is the relative frequency  $f_i/n$ , which is called weighting coefficient in general. The formula (1.2) is equivalent to the statement: the sample mean calculated based on a frequency table is a weighted mean of the mid-values with the frequencies as weighting coefficients.

### 1.3.2 Geometric mean

“Titer” is a widely applied measurement of concentration in microbiology and immunology where the tested material is proportionately diluted so that several samples with different concentrations are prepared and tittered respectively until certain phenomena appears, of which the corresponding diluted proportion is defined as the measurement of the concentration. For instance, The concentrations of certain antibody are measured for a set of sample and the corresponding titers are 4, 8, 16, 16, 64, 128, of which the arithmetic mean 39.3 is not an ideal representative of the data but the geometric mean is applied conventionally. The arithmetic mean of the logarithms of the titers is calculated firstly,

$$\frac{\log 4 + \log 8 + \log 16 + \log 16 + \log 64 + \log 128}{6} = 1.3045$$

then the anti-logarithm of it,  $\log^{-1} 1.3045 = 20.16$ , is the geometric mean of the above data set.

In general, if the individual values of the sample are all greater than 0, denoted by  $x_1, x_2, \dots, x_n$ , and the geometric mean is denoted by  $\bar{x}_g$ , then

$$\bar{x}_g = \log^{-1} \left( \frac{\log x_1 + \log x_2 + \dots + \log x_n}{n} \right) \quad (1.3)$$

or

$$\bar{x}_g = \sqrt[n]{x_1 x_2 \dots x_n}. \quad (1.4)$$

When the histogram of the sample is positive skew, if the histogram of the logarithms is close to symmetric, then the geometric mean may well represent the average level and usually it is less than the arithmetic mean.

### 1.3.3 Median

When the histogram of the sample shows taller around center and shorter on two sides but worse in symmetric, no matter positive skew or negative skew, the median, denoted by  $M_d$ , can be applied to measure the average level.

#### 1.3.3.1 Raw data based approach

Arrange the individual values in the sample from smallest to largest; when the number of individuals  $n$  is an odd number, the observed value with rank  $(n + 1)/2$  is taken as the median; when  $n$  is an even number, the average of the observed values with rank  $n/2$  and  $(n + 1)/2$  is taken as the median. For example, the median of the data set  $\{1, 1, 2, 2, 3, 4, 6, 9, 10\}$  is 3, while that of  $\{1, 1, 2, 2, 3, 4, 6, 9, 10, 13\}$  is  $(3 + 4)/2 = 3.5$ .

#### 1.3.3.2 Frequency table based approach

When the frequency table is available only, the median can be calculated approximately according to the following steps:

- (1) Calculate the rank corresponding to the median with  $n/2$  approximately (may not necessary be an integrate number).
- (2) Find out the sub-interval corresponding to the rank based on the cumulative frequencies, and denote with “ $a \sim b$ ”, of which the length is  $b - a$ .
- (3) Find out the cumulative frequencies up to the two ends of the sub-interval,

$f_a$  = the cumulative frequency of the last sub-interval,

$f_b$  = the cumulative frequency of the current sub-interval.

- (4) Estimate the value corresponding to the rank  $n/2$  through interpolation

$$M_d \approx a + \frac{b - a}{f_b - f_a} \left( \frac{n}{2} - f_a \right). \quad (1.5)$$

**Example 1.4** The two columns of Table 1.8 is the frequency table related to Fig. 1.4. Please calculate the arithmetic mean  $\bar{x}$ , geometric mean  $\bar{x}_g$  and median  $M_d$  of hair mercury for the residents of the city approximately on the basis of these data.

**Solution** The 4th column of Table 1.8 is that of mid-values. The individual values are approximately equal to these mid-values respectively, and hence

$$\begin{aligned}\bar{x} &\approx (20 \times 2 + 66 \times 4 + 60 \times 6 + 48 \times 8 + \cdots + 3 \times 20)/239 \\ &= 1598/239 = 6.69 \text{ } (\mu\text{mol/kg}) \\ \bar{x}_g &\approx \log^{-1}((20 \times \log 2 + 66 \times \log 4 + 60 \times \log 6 + 48 \\ &\quad \times \log 8 + \cdots + 3 \times \log 20)/239) \\ &= \log^{-1}(0.7711) = 5.90 \text{ } (\mu\text{mol/kg}).\end{aligned}$$

As for median, the corresponding rank is about

$$\frac{n}{2} = \frac{239}{2} = 119.5$$

which is located in the sub-interval “5~7”; the cumulative frequency up to “5” (the cumulative frequency of the sub-interval “3~5”) is 86; the cumulative frequency up to “7” (the cumulative frequency of the sub-interval “5~7”) is 146; through interpolation,

$$M_d \approx 5 + \frac{7 - 5}{146 - 86}(119.5 - 86) = 6.12 \text{ } (\mu\text{mol/kg}).$$

Table 1.8 The frequency table of hair mercury ( $\mu\text{mol/kg}$ ) for the residents of a city.

Sub-interval (1)	Frequency (2)	Cumulative frequency (3)	Mid-value ( $x$ ) (4)
1~	20	20	2
3~	66	86 ( $f_a$ )	4
5~ ( $a\sim b$ )	60	146 ( $f_b$ )	6
7~	48	194	8
9~	18	212	10
11~	16	228	12
13~	6	234	14
15~	1	235	16
17~	1	236	18
19~21	3	239	20
Total	239		

## 1.4 Measurement for Variation of a Sample

In addition to the measure for average level, the measure for variation among individual values is also necessary. The 4 measures frequently used are introduced as the follows.

### 1.4.1 Range $R$

It has been mentioned before that range is defined as the difference between the maximal value and the minimal value in the sample. Obviously, a bigger range indicates that the individual values are wider dispersed or higher varied. However, this measure depends on the maximal value and minimal value only while they often change a lot from sample to sample, and hence,  $R$  is worse in robustness.

### 1.4.2 $Q_3 - Q_1$

Arrange the  $n$  individual values in the sample from smallest to largest; the value with a rank mostly close to  $nP\%$  is called  $P\%$  quantile or  $P$  percentile of the sample, denoted by  $X_p$ . As special cases, 50% quantile or 50 percentile is exactly the median; 25% quantile or 25 percentile is called the first quartile, denoted by  $Q_1$ ; the 75% quantile or 75 percentile is called the third quartile, denoted by  $Q_3$ .

The difference between  $Q_3$  and  $Q_1$  is another measure for variation. A bigger  $Q_3 - Q_1$  indicates that the individual values are wider dispersed. Here the information on ranks of the data is partly used, hence the robustness of  $Q_3 - Q_1$  is better than that of range  $R$ .

The raw data based approach for  $P$  percentile is similar to that for median. Arrange the individual values in the sample from smallest to largest. If  $nP\%$  is an integer, then the value with this integer as rank is taken as the  $P$  percentile. Otherwise, there are two integers closing to  $nP\%$  and hence the average of two corresponding values is taken as the  $P$  percentile.

The steps of frequency table based approach for  $P$  percentile are also similar to that for median, only but  $n/2$  there should be changed with  $nP\%$ ,

$$X_p \approx a + \frac{b - a}{f_b - f_a} (nP\% - f_a). \quad (1.6)$$

### 1.4.3 Variance and standard deviation

Both of range and  $Q_3 - Q_1$  share the common shortcoming that the individual information cannot be used sufficiently and the inference on variation of the population can hardly be performed.

The difference between individual value and the population mean is called deviation from the mean. It could be positive or negative though its absolute value reflects the variation. The average of squared deviations throughout the population is called population variance, denoted by  $\sigma^2$ , of which the dimension is square of the variable's dimension. To make the dimension same as that of the variable, square root of the population variance is defined as population standard deviation, denoted by  $\sigma$ .

When the population mean is unknown and the sample data are available only, the population mean in the definition of deviation is replaced by the sample mean. It can be proved that the sum of squared deviations from the sample mean must be less than that of squared deviations from the population mean. To amend such a shortcoming, the sum is divided by  $(n - 1)$  instead of  $n$ , and hence the average sum of squared deviations is called sample variance, denoted by  $S^2$ ,

$$S^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1} \quad (1.7)$$

where  $(n - 1)$  is called degrees of freedom. In fact, since the restrain of

$$\sum_{i=1}^n (x_i - \bar{x}) = 0$$

among the  $n$  terms in the numerator of (1.7), only  $(n - 1)$  ones could vary freely.

For convenience in calculation, (1.7) can be expressed as

$$S^2 = \frac{\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2 / n}{n - 1}. \quad (1.8)$$

The readers can easily prove the equivalence between (1.7) and (1.8) with elementary algebra.

The square root of sample variance is called sample standard deviation, briefly denoted by  $S$  or  $SD$ , of which the dimension is the same as the variable itself. A bigger value of  $S$  refers to a higher variation.

### 1.4.4 Coefficient of variation

Sometimes the variations of two variables with different dimensions are needed to be compared. Obviously, their standard deviations cannot be compared directly because different dimensions. Then the coefficient of variation (CV), a concept without dimension, is useful, which is defined as

$$CV = \frac{S}{\bar{x}}. \quad (1.9)$$

Taking height and weight of normal young males as an example, assume the mean and standard deviation of height are 170cm and 6 cm and those of weight are 60kg and 7 kg; their standard deviations 6 cm and 7 kg are not comparable while the comparison between their coefficients of variation  $6/170 = 0.035$  and  $7/60 = 0.117$  shows that the variation of weight is higher than that of height.

Mean and standard deviation are two important numerical characters for describing continuous variables so that conventionally they are often expressed together as  $\bar{x} \pm s$ . For instance, the above mentioned mean and standard deviation of height could be expressed as  $170 \pm 6$  (cm), where the symbol “ $\pm$ ” just means “and” only.

## 1.5 Relative Measures and Standardization Approaches

### 1.5.1 Ratio, frequency and intensity

In vital statistics and epidemiology, relative measures are widely used to describe the probability and intensity of certain event happening to the individuals in the population and often named with “... rate”. However, with careful consideration one will find that there are three types of relative measures in fact.

#### 1.5.1.1 Ratio

It is simply a ratio of any quantity to another, such as

$$\text{Sex ratio of newly born babies} = \frac{\text{number of newly born girls}}{\text{number of newly born boys}}$$

and

$$\text{Mass index} = \frac{\text{Weight}}{\text{Height}^2} \text{ (kg/m}^2\text{)}$$

where the numerator and denominator may not necessary be counted numbers and nor with the same dimension.

### 1.5.1.2 *Relative frequency*

It is a special type of ratio where both of the numerator and denominator are counted numbers and the numerator is a part of the denominator. For a random sample, when the denominator is big enough, relative frequency approximately describes the chance of certain event happening to the individuals in the population. For example, if 90 patients were cured among 100 treated ones, then

$$\text{Cure rate} = \frac{\text{number of cured}}{\text{number of treated}} = \frac{90}{100} = 90\%.$$

There is no any dimension for relative frequency, and the value is a percentage or decimal within the interval of [0,1].

### 1.5.1.3 *Intensity*

It is another special type of ratio where the denominator is the total observed person years during certain period, the numerator is a number of certain event happening during the period. For example, the mortality rate is defined as

$$\left. \begin{array}{l} \text{Mortality} \\ \text{rate of} \\ \text{certain year} \end{array} \right\} = \frac{\text{Number of deaths during the year}}{\text{person years exposing to the risk of death during the year}}.$$

The dimension of numerator is “person”, that of denominator is “person × year” so that the dimension of mortality rate is “person/(person × year)” or “1/year”. If the denominator is regarded as the “adjusted total number of persons × 1 year”, then the mortality rate can be regarded as the adjusted relative frequency per year.

In general, intensity as a type of relative measures could be understood as “relative frequency per unit of time”, reflecting the chance of certain event happening in a unit of time.

If an inference for a relative measure from sample to population is needed, one has to recognize the type of it, whether it is simply a ratio or a relative frequency or an intensity, because different type requires different statistical method.

### 1.5.2 Crude death rate and standardization

We will use the mortality rate as an example to show why the crude intensities are not directly comparable and how the standardization approaches work.

Table 1.9 gives two sets of data for two cities respectively, each of which includes several age groups; for each age group, the mid-year population, number of deaths during the year and age specific mortality rate are available. Ignoring the age groups and dividing the total number of deaths by the sum of mid-year populations, the crude mortality rates can be calculated,  $P_a = 11.1\%$ ,  $P_b = 23.3\%$ . It seems that the risk of death in city B is higher than that in city A. However, in view of the age specific mortality rates, the risk of death in city A is higher than that in city B for all age groups. How to explain such a fallacy? Obviously, the crude mortality rate is incomparable because the age distributions are not balance between two cities; it is reasonable to compare the mortality rates age group by age group, but the variety of results based on separate comparisons can hardly be summarized into one conclusion.

Table 1.9 The data of age specific mortality rates for two cities.

Age group (year)	City A			City B		
	Mid-year population ( $10^3$ )	Number of deaths ( $10^3$ )	Mortality rate (%)	Mid-year population ( $10^3$ )	Number of deaths ( $10^3$ )	Mortality rate (%)
(1)	(2)	(3)	(4)	(5)	(6)	(7)
0~	400	2	5.0	288	1	3.5
15~	2000	10	5.0	238	1	4.2
30~	2000	15	7.5	794	5	6.3
45~	800	8	10.0	2000	18	9.0
60~	400	16	40.0	2000	70	35.0
75+	80	12	150.0	300	36	120.0
Total	5680	63	11.1	5618	131	23.3

A comprehensive measure summarizing all age specific mortality rates is often expected in applications such as comparison between different cities. There exist several methods for summarization sharing a similar idea — standardization, that is, to adjust the unbalance of age distributions by selecting certain “standard” and calculating standardized mortality rates.

### 1.5.2.1 Direct standardization approach

The main steps of direct standardization are the follows: select a “standard population” firstly; apply the whole set of age specific mortality rates to such a “standard population” and calculate the “expected number of deaths” for each age group in the “standard population”; calculate the crude mortality rate of the “standard population” based on the total expected numbers of deaths and call it a direct standardized mortality rate.

**Example 1.5** Taking the sum of populations of the two cities in Table 1.9 as a “standard population”, please compare the risk of death between two cities through the direct standardization approach.

**Solution** Column 2 of Table 1.10 refers to the standard population which consists the sum of the two populations for each age group; columns 3 and 5 refer to age specific mortality rates of two cities respectively; columns 4 and 6 refer to the expected numbers of deaths for each age group if the mortality rates were applied to the “standard population” correspondingly;

Table 1.10 The direct approach for standardized mortality rates of two cities.

Age group (year)	Standard population ( $10^3$ )	City A		City B	
		Mortality rate (%)	Expected number of deaths ( $10^3$ )	Mortality rate (%)	Expected number of deaths ( $10^3$ )
(1)	(2)	(3)	(4) = (2) × (3)	(5)	(6) = (2) × (5)
0~	686	5.0	3.43	3.5	2.40
15~	2238	5.0	11.19	4.2	9.40
30~	2794	7.5	20.96	6.3	17.60
45~	2800	10.0	28.00	9.0	25.20
60~	2400	40.0	96.00	35.0	84.00
75+	380	150.0	57.00	120.0	45.00
Total	11298	19.2	216.58	16.3	184.20

dividing the total expected numbers of death by the total population of the “standard population”, one can obtain the direct standardized mortality rates for the two cities and put in the bottom cells of columns 3 and 5 respectively; and it concludes that the standardized mortality rate of city A is higher than that of city B. This is consistent with the conclusion obtained by the comparisons age group by age group.

### 1.5.2.2 Indirect standardization approach

The main steps of indirect standardization are the follows: select a set of “age specific mortality rates” as the “standard” firstly; apply it to the studied population and calculate the “expected number of deaths” for each age group of it; calculate the ratio between the total observed number of deaths and the total expected number of deaths and call it standard mortality ratio (SMR); multiplying the crude mortality rate of the “standard” with SMR, one can obtain the indirect standardized mortality rate for the studied population.

**Example 1.6** Taking a set of age specific mortality rates as standard (see column 2 of Table 1.11), compare the risk of death between the cities A and B based on the data in Table 1.9 through the indirect standardization approach.

**Solution** Columns 3 and 5 of Table 1.11 refer to the studied populations of the two cities; columns 4 and 6 refer to the expected numbers of deaths if the standard age specific mortality rates were applied to the studied populations respectively; dividing the total observed numbers of death (see columns 3 and 6 in Table 1.9) by the total expected numbers of deaths (see Table 1.11), one can obtain the SMRs for the two cities; multiplying the crude mortality rate of the “standard” with SMRs, one can obtain the indirect standardized mortality rates for the cities A and B respectively.

$$\text{City A: SMR} = 63/58.12 = 1.084$$

$$\text{Indirect standardized mortality rate} = 17.2 \times 1.084 = 18.64(\%).$$

$$\text{City B: SMR} = 131/142.30 = 0.920$$

$$\text{Indirect standardized mortality rate} = 17.2 \times 0.920 = 15.83(\%).$$

Table 1.11 The indirect approach for standardized mortality rates of two cities.

Age group (year)	Standard mortality rate (%)	City A		City B	
		Mid-year population ( $10^3$ )	Expected number of deaths ( $10^3$ )	Mid-year population ( $10^3$ )	Expected number of deaths ( $10^3$ )
(1)	(2)	(3)	(4) = (2) × (3)	(5)	(6) = (2) × (5)
0~	4.3	400	1.72	288	1.23
15~	4.6	2000	9.20	238	1.08
30~	6.9	2000	13.80	794	5.43
45~	9.5	800	7.60	2000	19.00
60~	37.5	400	15.00	2000	75.00
75+	135.0	80	10.80	300	40.50
Total	17.2	5680	58.12	5620	142.30

Comparing the SMRs or the indirect standardized mortality rates between the two cities, one can find that the risk of death in the city A is more serious than that in the city B.

### 1.5.2.3 Nature of crude mortality rate and standardized mortality rate

The crude mortality rate is a weighted average of age specific mortality rates with the sub-populations of age groups as the weight coefficients. If there are higher age specific mortality rates in the age groups with more populations, then the crude mortality rate is higher. Table 1.9 shows that the structures of populations in the two cities are obviously different, that is, more youngling in city A and more elderly in city B. Therefore, offering higher weights to the higher age specific mortality rates, the weighted average results in a higher crude mortality rate of city B than that of city A.

In order to solve the problem of unequal weight coefficients, the idea of weighted average is still used in the direct standardization approach, but where the sub-populations of age groups in the “standard population” are taken as the weight coefficients. Sometimes, different standard populations selected might result in quite different direct standardization mortality rates.

Totally giving up the information on age specific mortality rates, the indirect standardization approach keeps that on the numbers of death only. In

fact, it is to calculate a weighted average of the selected standard age specific mortality rates with the observed sub-populations as the weight coefficients firstly; then SMR and then use it to magnify or dwindle the weighted average. Similarly, different sets of standard age specific mortality rates selected might result in quite different indirect standardization mortality rates.

The selection of standard populations or standard mortality rates is fairly important. Usually the populations or mortality rates of the world or the country or the province are considered as the standard. If it is intended to compare two cities only, then the pool of the two populations or the pooled estimation of the age specific mortality rates (sum of the numbers of deaths in the age group/the sum of the sub-populations) might be taken as the standard. In practice, it is considerable to select more than one standard to see whether the results are consistent or not. If it is consistent, then the conclusion might be reliable; otherwise, one should be careful.

## 1.6 Computer Experiments

**Experiment 1.1 Frequency table and histogram** The detailed steps in the software for frequency table (Program 1.1) are similar to that in hand operation. Assume that the data in Example 1.4 have been input into an ASCII coded file “RBC.DAT”, now the software (such as *Edit*) is needed to perform a frequency table with the following steps:

**Step 1** Find out the minimum and maximum

Lines 01–09, read the data and find out the minimum and maximum.

**Step 2** Design the sub-groups

By calculating the range and deciding the number of sub-groups, it is obtained as the follows:

Sub-group	Mid-value
3.20~	3.35
3.50~	3.65
3.80~	3.95
4.10~	4.25
4.40~	4.55
4.70~	4.85
5.00~	5.15
5.30~	5.45
5.60~	5.75
5.90~6.20	6.05

Program 1.1 Frequency table and histogram.

Line	Program	Line	Program
01	DATA RBC;	15	IF X<4.40 & X>=4.10 THEN Y=4.25;
02	INPUT X @@;	16	IF X<4.70 & X>=4.40 THEN Y=4.55;
03	CARDS;	17	IF X<5.00 & X>=4.70 THEN Y=4.85;
04	5.12 5.13	18	IF X<5.30 & X>=5.00 THEN Y=5.15;
05	.....	19	IF X<5.60 & X>=5.30 THEN Y=5.45;
06	3.86 5.69	20	IF X<5.90 & X>=5.60 THEN Y=5.75;
07	;	21	IF X>=5.90 THEN Y=6.05;
08	PROC MEANS MIN MAX;	22	PROC UNIVARIATE FREQ;
09	RUN;	23	VAR Y;
10	DATA FRBC;	24	RUN;
11	SET RBC;	25	PROC GCHART;
12	IF X<3.50 THEN Y=3.35;	26	VBAR Y/TYPE=PERCENT;
13	IF X<3.80 & X>=3.50 THEN Y=3.65;	27	VBAR Y/TYPE=CPERCENT;
14	IF X<4.10 & X>=3.80 THEN Y=3.95;	28	RUN;

**Step 3** Organize data and list frequency table

Lines 10–21, each value is changed with the corresponding mid-value of its sub-group; lines 22–24, calculate description statistics such as mean, variance, standard deviation and variation coefficient (although the median and quantile could be given namely, they are just the mid-values in their sub-intervals instead of the values obtained by interpolation introduced above) and the frequency table is performed.

**Step 4** Histogram

Lines 25–28, work out the histograms for frequency and cumulative frequency.

**Experiment 1.2 Calculation of standardized mortality rate with direct approach and indirect approach** Program 1.2 is the SAS program for reference. The first 20 lines are for the direct approach where lines 4–5

Program 1.2 Program for direct approach and indirect approach.

Line	Program	Line	Program
01	DATA STA;	24	A1=P1*SP/1000;
02	INPUT P1 D1 P2 D2;	25	A2=P2*SP/1000;
03	KEEP SP P1 R1 A1 A2;	26	CARDS;
04	R1=D1/P1*1000;	27	4.3 400 286
05	R2=D2/P2*1000;	28	4.6 2000 238
06	SP=P1+P2;	29	6.9 2000 794
07	A1=R1*SP/11298;	30	9.5 800 2000
08	A2=R2*SP/11298;	31	37.5 400 2000
09	CARDS;	32	135.0 80 300
10	400 2 286 1	33	;
11	2000 10 238 1	34	PROC PRINT;
12	2000 15 794 5	35	PROC MEANS SUM NOPRINT;
13	800 8 2000 18	36	VAR A1 A2;
14	400 16 2000 70	37	OUTPUT OUT=STAN3 SUM=STA STB;
15	80 12 300 36	38	DATA STAN4;
16	;	39	SET STAN3;
17	PROC PRINT;	40	KEEP STA STB SMRA SMRB SMPA SMPB;
18	PROC MEANS SUM;	41	SMRA=63/STA;
19	VAR A1 A2;	42	SMRB=131/STB;
20	RUN;	43	SMPA=SMRA*17.2;
21	DATA STA2;	44	SMPB=SMRB*17.2;
22	INPUT SP P1 P2;	45	PROC PRINT;
23	KEEP SP P1 P2 A1 A2;	46	RUN;

calculate the age specific mortality rates, lines 7–8 calculate the age specific numbers of deaths, lines 10–17 list the data, and lines 18–19 calculate the standardized mortality rate.

Lines 21–46 are for the indirect approach where standardized mortality rates and sub-populations are required as input. Lines 24–25 calculate age specific expected numbers of deaths; lines 27–34 list the data; lines 35–37 calculate two total expected numbers of deaths respectively and put into STAN#; lines 41–44 calculate SMR and the standardized mortality rate respectively; then line 45 prints out the results.

## 1.7 Thinking, Practice and Experiments

1. True or false. Which of the following statements are correct and which wrong?

- (1) “Whether there are red cells in occult blood examination” is a continuous variable.
- (2) Red cell count is a discrete variable.
- (3) The arithmetic mean is always greater than the median.
- (4) The mean of large sample is always more closer to the population mean than that of small sample.
- (5) The arithmetic mean is always greater than the standard deviation.
- (6) A histogram can be used to describe the distribution of the weight of a group of newborn babies.
- (7) The distribution of the days of hospitalization for certain disease shows higher around center and lower on two sides; the arithmetic mean is 10 days and the median is 5 days. One can see that the distribution is positive skew.
- (8) The dimension of variation of coefficient is the same as that of the original variable.
- (9) If the sample mean is greater, then the standard deviation must be greater.
- (10) The range may increase with the increase of sample size.

2. Calculate the sample mean, median, variance, standard deviation and coefficient of variation for Example 1.4 on the basis of the raw data and the frequency table respectively; then compare and discuss the two sets of results.

3. The blood-glucose (mmol/L) is measured for 12 randomly selected patients. The data are 5.31, 6.12, 6.53, 6.53, 6.65, 6.66, 6.71, 6.93, 7.05, 7.15, 7.21, 7.35. Please calculate the arithmetic mean, geometric mean and median; which one better reflects the average level? Again calculate the range,  $Q_3 - Q_1$  and standard deviation; which one better reflects the variation?

4. The daily fat intake (g) of 100 randomly selected adults was surveyed with the data as the follows:

23	60	78	84	90	104	114	127	130	143
43	69	81	94	97	102	117	120	147	150
52	80	88	96	103	105	114	128	130	153
65	79	89	95	107	108	128	131	139	148
67	75	76	91	102	105	127	138	153	167
70	72	95	103	111	117	128	130	147	142
67	62	72	95	109	111	127	132	144	151
23	37	69	88	99	109	119	139	134	155
30	89	76	96	93	104	117	133	147	151
44	73	83	94	96	107	111	128	131	150

Please work out a frequency table and a histogram; calculate the arithmetic mean, variance, standard deviation and coefficient of variation as well as median and  $Q_3 - Q_1$ .

5. Calculate the approximate arithmetic mean and standard deviation of red cell counts of 120 normal male adults based on the frequency table (Table 1.6) and compare with those calculated based on the raw data. Through this example, can you summarize the main steps for calculating arithmetic mean and standard deviation based on a frequency table in general?

6. It is quite popular to use two different concepts to describe the incidence of disease:

$$\left. \begin{array}{l} \text{Cumulative} \\ \text{incidence} \\ \text{rate} \end{array} \right\} = \frac{\text{Number of new patients during the same period}}{\text{Total number of persons followed during certain period}}.$$

$$\left. \begin{array}{l} \text{Person-year} \\ \text{incidence} \\ \text{rate} \end{array} \right\} = \frac{\text{Number of new patients during the same period}}{\text{Total person-years exposed during certain period}}.$$

Please discuss the properties of these two rates; are they ratio, frequency or intensity?

7. The data of liver-cancer specific mortality rates for males in two cities are collected as the follows (Gong Zhiping, 1992):

Age group (year)	City A				City B			
	Population	Proportion	Number of deaths	Mortality rate	Population	Proportion	Number of deaths	Mortality rate
0~	323600	0.6555	24	7.4	364500	0.6949	22	6.0
30~	56800	0.1150	75	132.0	64300	0.1226	75	116.6
40~	42400	0.0850	103	242.9	40100	0.0765	104	259.4
50~	30500	0.0618	87	285.2	28800	0.0549	84	291.7
60~	21300	0.0431	69	323.9	16200	0.0309	54	333.3
70~	19100	0.0387	33	172.8	10600	0.0202	22	207.5
Total	493700	1.0000	391	79.2	524500	1.0000	361	68.8

Please compare the risk of liver cancer to the males between the two cities through the direct standardization approach.

- (1) Taking the population of city A as a standard population;
- (2) Taking the population of city B as a standard population;
- (3) Taking the total population of cities A and B as a standard population;
- (4) Compare and discuss the results.

8. Please compare the risk of liver cancer to the males between the two cities through the indirect standardization approach.

- (1) Taking the age specific mortality rates of city A as standard mortality rates;
- (2) Taking the age specific mortality rates of city B as standard mortality rates;
- (3) Taking the pooled age specific mortality rates of city A and B as standard mortality rates;
- (4) Compare and discuss the results.

9. Prove or check the following statements. Assume there are observed values  $y_1, y_2, \dots, y_n$ , and denote  $\bar{y} = \sum_{i=1}^n (y_i/n)$ .

$$(1) \sum_{i=1}^n ay_i = a \sum_{i=1}^n y_i;$$

$$(2) \sum_{i=1}^n (y_i - \bar{y}) = 0;$$

$$(3) \sum_{i=1}^n (a + y_i) = na + \sum_{i=1}^n y_i;$$

$$(4) \sum_{i=1}^n \left(\frac{y_i}{n} a\right) = a \sum_{i=1}^n \frac{y_i}{n};$$

$$(5) \sum_{i=1}^n (y_i + a)^2 = \sum_{i=1}^n y_i^2 + 2a \sum_{i=1}^n y_i + na^2;$$

$$(6) \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - \frac{1}{n} \left(\sum_{i=1}^n y_i\right)^2;$$

$$(7) \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - n\bar{y}^2.$$

(1st edn., Ji-Qian Fang, Qing Liu; 2nd edn., Jiansheng Ding, Futian Luo, Aihua Lin; 3rd edn., Ji-Qian Fang.)