

Chapter 1

Descriptive Statistics

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 block plot, histogram) and numerical indices (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 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 the 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, salesman 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 the simplest special case of it.

The number of individuals within a 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 in 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, un-effective 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.

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

No.	Age (years)	Gender	Treatment	Systolic pressure (kPa)	Diastolic pressure (kPa)	ECG	Effectiveness
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

1.1.2 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 the statistical software such as SAS, SPSS, etc.

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 after treatment, then a patient is defined as an observed unit; otherwise, if the systolic pressure and diastolic pressure are measured at 3 different times 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. In Table 1.1, columns 2–8 show a 100×7 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.

1.2 Frequency Table and Histogram

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

1.2.1 Frequency table

As mentioned before, in a set of samples, the number of times a certain event occurs 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

For 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 1.2, two frequency tables are given in Tables 1.2 and 1.3,

Table 1.2 The frequency table for gender of 108 patients.

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.

Occupation	Frequency	Relative frequency (%)	Cumulative frequency	Cumulative relative frequency (%)
Worker	28	25.9	28	25.9
Farmer	23	21.3	51	47.2
Businessman	24	22.2	75	69.4
Student	18	16.7	93	86.1
Soldier	15	13.9	108	100.0
Total	108	100.0		

where the ratio between the frequency and the total number is called relative frequency (if no confusion will arise, it is also called frequency). The sum of all relative frequencies must be 100% (in practice, sometimes it is not exactly 100% due to rounding error). The cumulative frequencies and cumulative relative frequencies are the results of 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.

1.2.1.2 Continuous type frequency table

For continuous variable, the general method to establish a frequency table could be learnt from the following example.

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

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.

Effectiveness	Frequency	Relative frequency (%)	Cumulative frequency	Cumulative relative frequency (%)
Cure	65	43.3	65	43.3
Effect	45	30.0	110	73.3
No effect	25	16.7	135	90.0
Worse	15	10.0	150	100.0
Total	150	100.0		

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

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 the 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 first, and then slightly adjust to an 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 the fact 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).
- (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).

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

Sub-interval	Mark	Frequency	Relative frequency (%)	Cumulative frequency	Cumulative relative frequency (%)
3.20–	┐	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┐	19	15.8	36	30.0
4.40–	正正正正┐	23	19.2	59	49.2
4.70–	正正正正┐	24	20.0	83	69.2
5.00–	正正正正┐	21	17.5	104	86.7
5.30–	正正┐	11	9.2	115	95.9
5.60–	┐	4	3.3	119	99.2
5.90–6.20	┐	1	0.8	120	100.0
Total		120	100.0		

- (5) Calculate the frequencies, relative frequencies and cumulative frequencies (columns 3–6 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 axis refers to “various situations” of the variable and the vertical axis refers to the corresponding frequencies.

1.2.2.1 Frequency plot for discrete variable — bar chart

For a discrete variable, one can use the points on the horizontal axis 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.

1.2.2.2 Frequency plot for continuous variable — histogram

For a continuous variable, one can use the sub-intervals with equal length on the horizontal axis to express the different situations of the variable; and

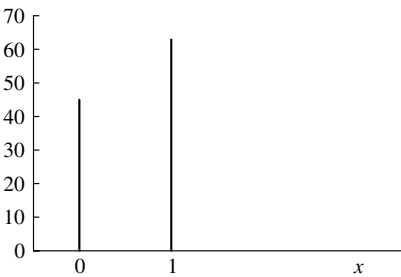


Fig. 1.1 The frequency plot for gender of 108 patients. x : gender, 0: female, 1: male.

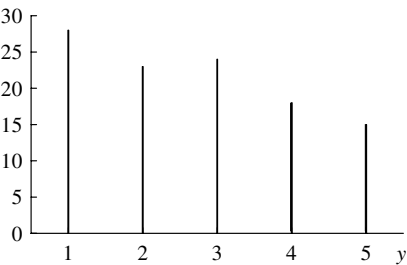


Fig. 1.2 The frequency plot for occupation of 108 patients. y : occupation, 1: worker, 2: farmer, 3: businessman, 4: student, 5: soldier.

plot vertical rectangles on these intervals, of which the heights express the frequencies related to the sub-intervals (Fig. 1.3(a)), this is called histogram. 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. The height of any rectangle in a histogram is neither the frequency nor the relative 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 regardless of the lengths of the sub-intervals.

Both 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 the center

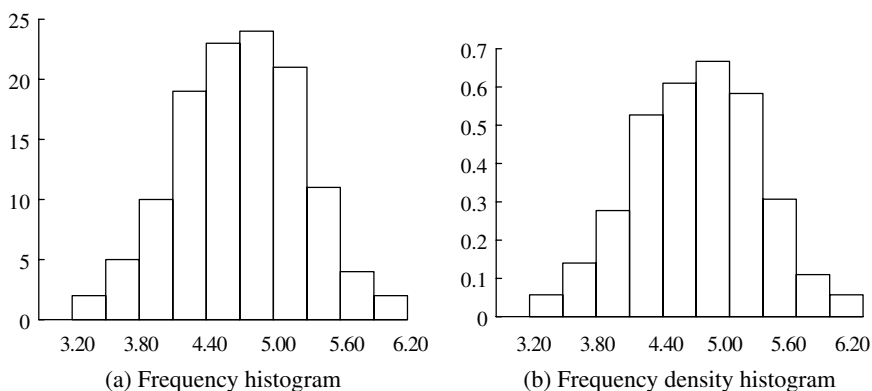


Fig. 1.3 Histograms plotted on the basis of the frequency table for the data set of red blood cell counts ($10^{12}/L$) of 120 normal male adults.

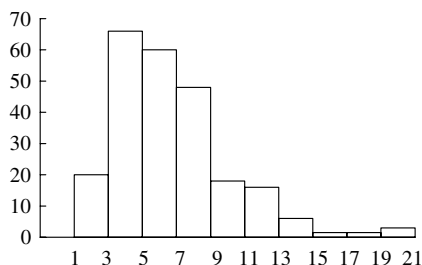


Fig. 1.4 Frequency histogram of hair mercury for the residents of a city.

and shorter on two sides, which indicate that the red blood 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. 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 the 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 evaluating the quality of life is given in Fig. 1.6. One can see that 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

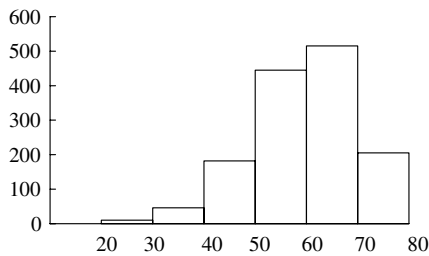


Fig. 1.5 Frequency histogram of age for a group of male lung cancer patients.

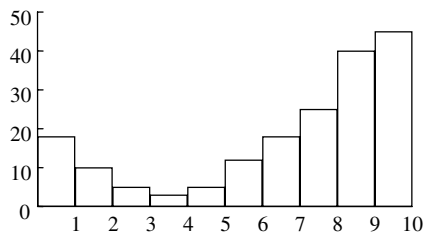


Fig. 1.6 Frequency distribution of the satisfactory score to an exhibition among the visitors.

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 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 not 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 four categories: cure, effect, no effect and worse, and the corresponding frequencies can be expressed with four bars on the horizontal axis as a bar chart for discrete variable.

1.2.3 Cumulative frequency plot

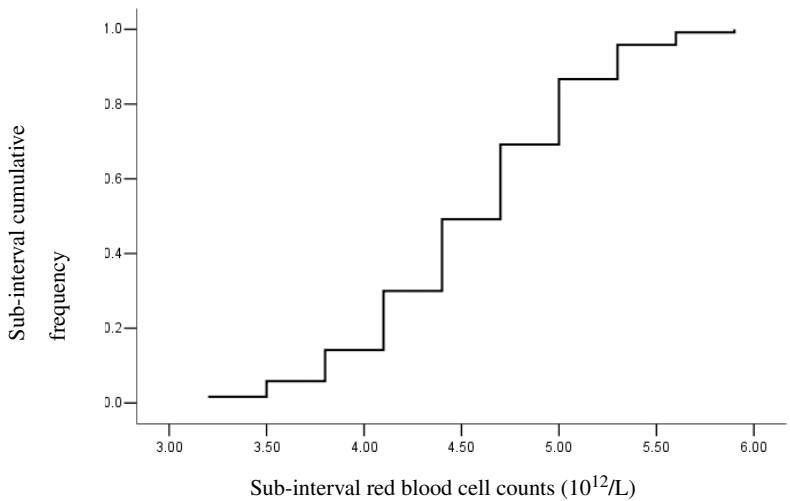
We can also use cumulative frequency plot to show how the frequency and percentage of individuals accumulate as the value increases, where

the horizontal axis refers to “various situations” of the variable and the vertical axis refers to the cumulative frequencies. According to Table 1.6, column 5 indicates the cumulative percentages at each observed red blood cell counts level among 120 normal male adults. We can get the cumulative frequency distribution based on the frequency table when using the data in column 5 as the values for vertical axes, and the upper limit values in column 1 as the values for horizontal axes (Fig. 1.7(a)). We can also get the cumulative frequency distribution based on the raw data (Fig. 1.7(b)). For example, there were 120 observations in Example 1.3, so each represents $1/120 = 0.83\%$. The first observation ($3.29 \times 10^{12}/L$) corresponds to a cumulative frequency of 0.83%, the first and second observations to a cumulative frequency of 1.67%, and so on. Cumulative frequency distribution is useful in finding the median and other quartiles. We can easily get the median, the lower and upper quartiles (25% and 75% quartiles) according to Fig. 1.7(b). The cumulative frequency distribution is a continuous ladder shape curve, say, the vertical jumps correspond to the increases in the cumulative frequencies at each observed red blood cell counts level. The cumulative frequency curve is steep when there is a concentration of values, and shallow when the values are sparse. In Fig. 1.7, the curve is steep in the center, and shallow around the low and high values. This means the majority of red blood cell counts are concentrated in the center of the distribution.

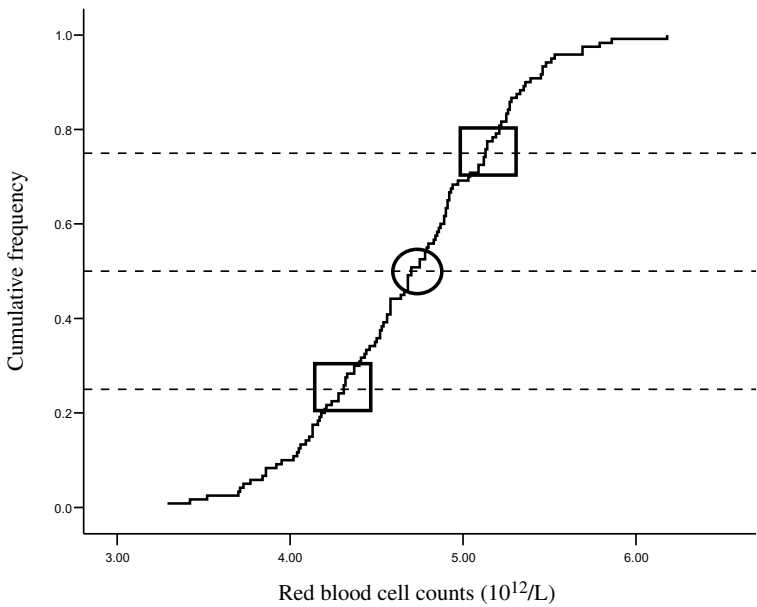
We usually use histograms to compare the distributions of two variables. However, cumulative frequency distribution provides more information when the histograms overlap extensively. For example, the histograms of the outcomes corresponding to the new medication group and control group mostly overlap in Fig. 1.8, thus we can hardly describe the difference between the two groups simply based on the histograms. In contrast, viewing at the cumulative frequency distribution, we can find that the change is sharper in the control group than that in the new medication group.

1.3 Measurement for Average Level of a Sample

In addition to frequency table and histogram, several numerical characteristics are also used for statistical description. For continuous variables, two often used characteristics are the average level and variation.



(a)



(b)

Fig. 1.7 Cumulative frequency distribution of the red blood cell counts ($10^{12}/L$) of 120 normal male adults.

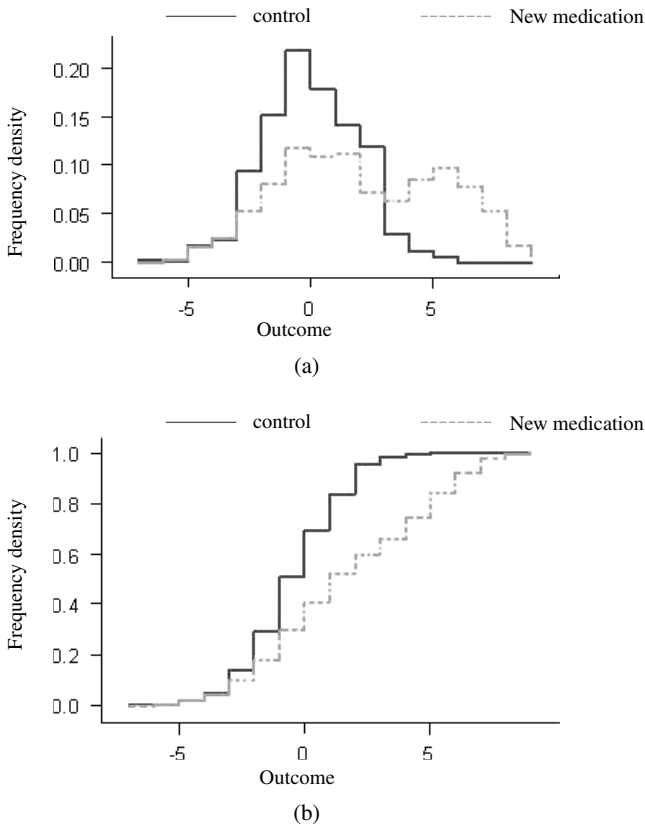


Fig. 1.8 Histograms and cumulative frequency distribution of effects outcome of the new medication group and the control group.

1.3.1 Arithmetic mean

When the histogram looks symmetric, the value that 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

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

Whenever no confusion arises, $\sum_{i=1}^n x_i$ could be simplified as $\sum_i x_i$ or even $\sum x$. Equation (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 from Table 1.7, the mean is

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

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

The formula for the above approach can be expressed as

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

where f_i and x_i are the frequency and the mid-value of the i th sub-interval, n is the total sample size. One can see from the process of the 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	0.8736
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	0.5014
5.60–	5.75	4	0.033	0.1897
5.90–6.20	6.05	1	0.008	0.0484
Total		120	1	4.6939

calculation that 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 relative frequency f_i/n in (1.2) that reflects the importance of the mid-value x_i 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 titered respectively until certain phenomenon 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, and 128, of which the arithmetic mean 39.3 is not an ideal representative of the data but the geometric mean. The arithmetic mean of the logarithms of the titers is calculated firstly,

$$(\log 4 + \log 8 + \log 16 + \log 16 + \log 64 + \log 128)/6 = 1.3045$$

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 with x_1, x_2, \dots, x_n , and the geometric mean is denoted with \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 it is usually less than the arithmetic mean.

1.3.3 Median

When the histogram of the sample is taller around center and shorter on two sides but worse in symmetry, no matter positive skew or negative skew, the median, denoted with 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 $(n/2)+1$ 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 only the frequency table is available, 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 necessarily be an integer);
- (2) Find out the sub-interval corresponding to the rank based on the cumulated frequencies, and denote with “ $a \sim b$ ” of which the length is $b - a$;
- (3) Find 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

$$M_d \approx a + \frac{b - a}{f_b - f_a}(0.5n - f_a) \quad (1.5)$$

Example 1.4 The two columns of Table 1.8 is the frequency table related to Fig. 1.4. Calculate the arithmetic mean \bar{x} , geometric mean \bar{x}_g and median

Table 1.8 The frequency table of hair mercury (μ mol/kg) for the residents of a city.

Sub-interval	Frequency	Cumulative frequency	Mid-value (x)
1–	20	20	2
3–	66	86 (f_a)	4
5– ($a-b$)	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		

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

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

As for median, the corresponding rank is about

$$n/2 = 239/2 = 119.5$$

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

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

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 four measures frequently used are introduced as 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 but 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 the smallest to largest; the value with a rank mostly close to $nP\%$ is called $P\%$ quartile or P percentile of the sample, denoted with X_p . As special cases, 50% quartile or 50th percentile is exactly the median; 25% quartile or 25th percentile is called the lower quartile, denoted with QL ; the 75% quartile or 75th percentile is called the upper quartile, denoted with QU .

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

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

The steps of frequency table based approach for P th percentile are also similar to those for median, only that $n/2$ 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 the 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 the population variance, denoted by σ^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 the population standard deviation, denoted with σ .

When the population mean is unknown and only the sample data are available, the population mean in the definition of deviation is replaced by the sample mean. It can be proved that the sum of the squared deviations from the sample mean must be less than that of the 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 the sample variance, denoted with S^2 ,

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

where $n - 1$ is called the 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), there are only $n - 1$ deviations which could be varied freely.

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

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

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

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

1.4.4 Coefficient of variation

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

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

Taking the height and weight of normal young males as an example, assume the mean and standard deviation of the height are 170 cm and 6 cm and those of the weight are 60 kg 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 greater 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 the variable of height could be expressed as 170 ± 6 (cm), where the symbol “ \pm ” just means “and”.

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 in fact three types of relative measures.

1.5.1.1 Ratio

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

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

and

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

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

1.5.1.2 *Relative frequency*

It is a special type of ratio where both the numerator and denominator are counted numbers and the numerator is part of the denominator. For a random sample, when the denominator is big enough, a 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 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

$$\begin{aligned} & \text{Mortality rate of certain year} \\ &= \frac{\text{number of deaths during the year}}{\text{person-years exposure to the risk of death during the year}}. \end{aligned}$$

The dimension of numerator is “person”, that of denominator is “person \times year” so that the dimension of mortality rate is “person/(person \times year)” or “1/year”. If the denominator is regarded as the “adjusted total number of persons \times 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 distributions of age are not balanced between the two cities; it

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 (%)
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	5620	131	23.3

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.

A comprehensive measure summarizing the comparison between two sets of age specific mortality rates is often expected in applications such as comparison between different cities. There exist several methods for summary sharing a similar idea — standardization, that is, to adjust the imbalance in 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 as 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”, compare the risk of death between the two cities through the direct standardization approach.

Solution Column 2 of Table 1.10 refers to the standard population which is the sum of the two populations for each age group; columns 3 and 5 refer

Table 1.10 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	45.60
Total	11298	19.2	216.58	16.3	184.20

to the age specific mortality rates of the two cities respectively; columns 4 and 6 refer to the expected number of deaths for each age group if the mortality rate were applied to the “standard population” correspondingly; dividing the total expected numbers of deaths by 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 age group comparison.

1.5.2.2 Indirect standardization approach

The main steps of indirect standardization are as follows: Select a set of “age specific mortality rates” as the “standard” first, 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 numbers 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 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 deaths (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 cities *A* and *B* respectively.

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 of City A (10^3)	Expected number of deaths in A (10^3)	Mid-year population of City B (10^3)	Expected number of deaths in B (10^3)
(1)	(2)	(3)	(4)=(2)×(3)	(5)	(6)=(2)×(5)
0–	4.3	400	1.72	288	1.24
15–	4.6	2000	9.20	238	1.09
30–	6.9	2000	13.80	794	5.48
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	142.31

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

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

$$\text{City B: SMR} = 131/142.31 = 0.921$$

$$\text{Indirect standardized mortality rate} = 17.2 \times 0.921 = 15.84(\%)$$

Comparing the SMRs or the indirect standardized mortality rates between the two cities, one can find that the risk of death in city A is much higher 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 youths in city A but 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 weights, 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 weights. 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 deaths 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 weights first; then SMR and use it to magnify or dwindle on 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 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 desirable 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 Frequently Used Graphs in Statistics

The first step of analysis is often to summarize and display the data, which can help us to identify outliers and possible errors in the data. Statistical chart is the important tool to display the data, which is intuitively clear by using the point-line-plane. There are several frequently used graphs in statistics, such as bar chart, percent bar chart, pie chart, line chart, semi-logarithmic line chart, box plot, and stem-and-leaf plot.

1.6.1 Layout of graphs

We can use Fig. 1.9 as an example to illustrate the layout of statistical graphs. The whole area for a graph is the chart area; the area within the X -axis and Y -axis is the drawing area; points and lines represent the original data. A two-dimensional graph consists of a horizontal coordinate axis (X -axis) and a vertical coordinate axis (Y -axis). For a three-dimensional graph, there is a third coordinate axis named Z -axis. There are scales on the

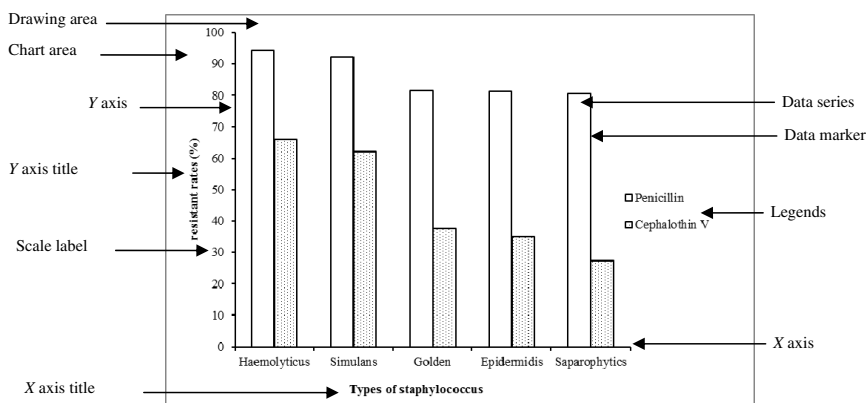


Fig. 1.9 Resistance rates of five types of staphylococcus for two kinds of antibiotics.

coordinate axes, and the corresponding numbers on the scales are named scale labels which could be real numbers or categories. Axes titles are left-aligned along the Y-axis and Z-axis or below the X-axis. There is no axes title for the pie chart. The basic rules for the layout of graphs are as follows:

- (1) Make sure that the graph is appropriate for the data and to support the main purpose of the study;
- (2) The title should be at the bottom of the graph;
- (3) Use different colors or different patterns for the different themes in the graphs, and put the legends at the appropriate place (at the right of the graph, bottom of the graph or top of the graph);
- (4) For the graphs which contain coordinate axes (bar chart, line chart, etc.), the values assigned for the X-axis should be in ascending order from the left to the right, and the values assigned for the Y-axis should be in ascending order from the bottom to the top. For the numerical variables, origin of coordinates, units and the appropriate scales should be labeled; for the categorical variables, the categories should be labeled. To make the graphs clearer, the height–width ratio is usually 5:7 (so-called “golden proportion”).

There are several kinds of software to create the statistical graphs, such as Excel, SAS, SPSS, R, Maple, Matlab. We will list the SAS program for Fig. 1.9 in Sec. 1.7.

1.6.2 Several graphs in statistics

Different situations call for different types of graphs, and it helps to have a good knowledge of what graphs are available.

- (1) Bar chart or bar diagram: In a bar chart the heights of the bars are drawn for the frequencies for each category of a set of data. There are two kinds of bar chart, simple bar chart (Fig. 1.10) and clustered bar chart (Fig. 1.9). The axis of the heights (usually the vertical axis) must begin at 0 (Fig. 1.11), or it may mislead the truth from the graphs. For example, in Fig. 1.10, if the vertical axis starts at 2, the proportional relationship visually appears to be $A : B = 2 : 1$, which disguises the fact that the proportional relationship of A and B is $4 : 3$. Each bar is in descending order of the variable in order to compare, and the space between two bars needs to be appropriate with a clear appearance.
- (2) Percent bar chart: The percent bar chart is used to display the frequency distribution. For example, Fig. 1.12 is plotted with the data in Table 1.12, where two bars with length equal to 100% are drawn for the two categories (hospitalization ≤ 7 days and > 7 days) at first

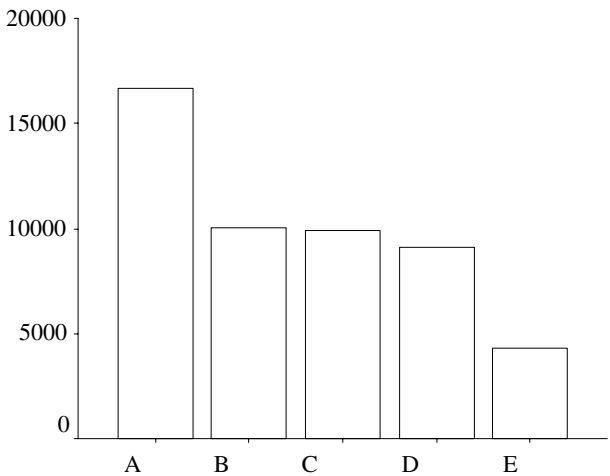


Fig. 1.10 Outpatient amount of the department of general internal medicine in the affiliated hospital of one medical university.
* A = Digestive; B = Cardiovascular; C = Respiratory; D = Endocrinology; E = Hematology.

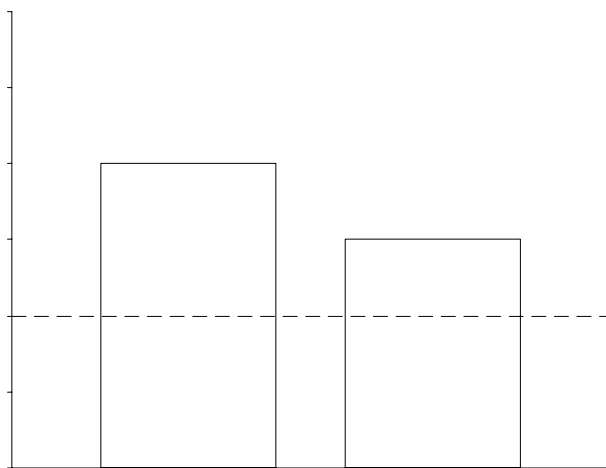


Fig. 1.11 The vertical axis must start at 0 in the bar chart.

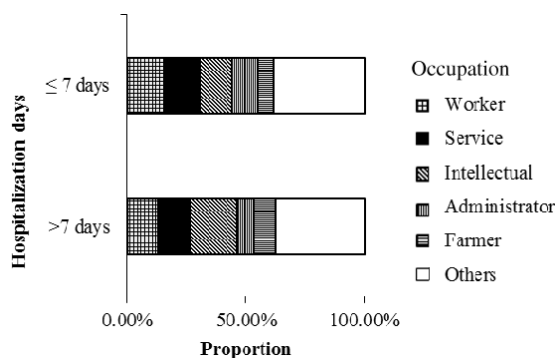


Fig. 1.12 The frequency distribution of the occupation of maternity patients with different hospitalization days.

step; and they are divided into several parts according to the proportion of the percentages of their component respectively in the second step. The sub-divided parts are sorted according to professional knowledge. If there is no exact sorting orders based on the professional knowledge, then they are usually in descending order by the proportion. The “other” is usually placed at the end of the bar.

- (3) Pie chart: The situation and the sorting orders for the pie chart are the same as the percent bar chart. The whole area (and consequently its

Table 1.12 The frequency distributions of occupations of maternity inpatients.

Occupation	Hospitalization ≤ 7 days	Hospitalization > 7 days
Worker	15.31	13.09
Farmer	6.79	9.06
Intellectual	13.40	19.46
Administrator	10.78	7.38
Service	15.22	13.42
Others	38.50	37.59

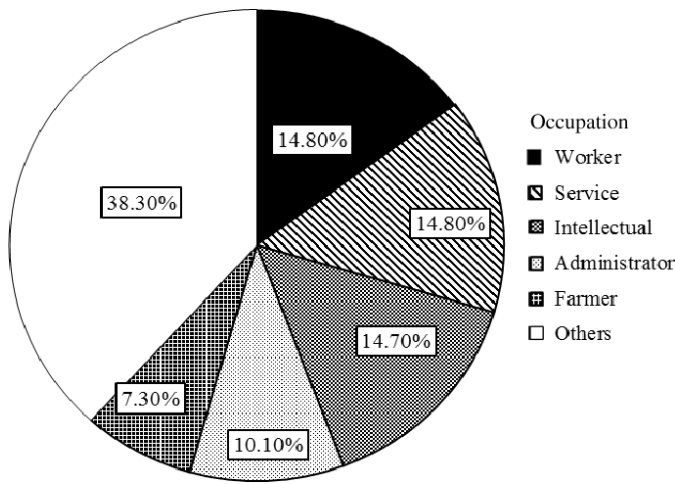


Fig. 1.13 The frequency distribution of the occupation of maternity patients.

central angle) of the pie equal to 100%, and the circle is divided into sectors to illustrate the according proportions. The proportion of the first sector starts at the 12 o'clock position. If there is no professional concern, it is usually sorted from large to small value. Figure 1.13 uses the data of Table 1.13. It indicates the occupation distribution among 1402 maternity patients. It is obvious that the percent bar chart is better than pie chart when comparing proportions between multiple sets of data.

- (4) Line chart: The lines up and down in the rectangular plane coordinate system are used to display trends over time, or the changing process

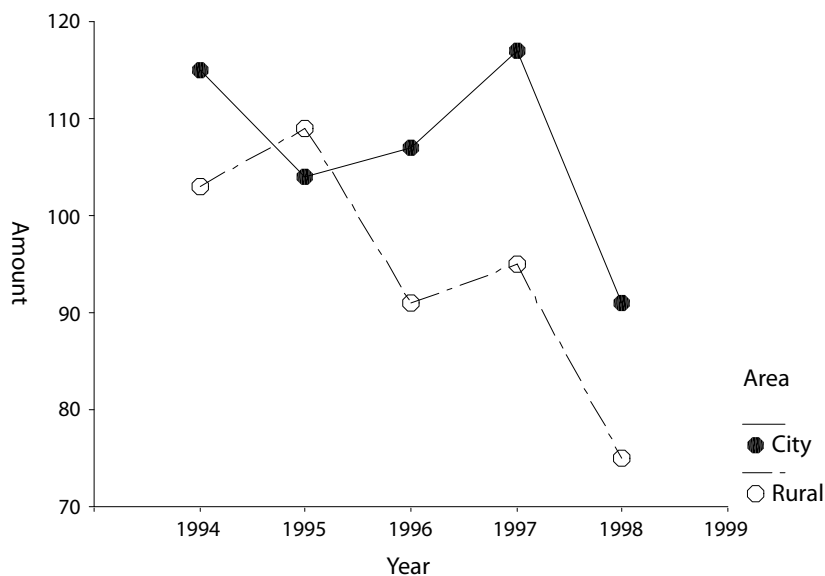


Fig. 1.14 The amount of discharged patients of the department of stomatology in the affiliated hospital of one medical university from 1994 to 1998.

Table 1.13 The mortalities of diarrhea and whooping cough (1/million) (1975–2000).

Year	Diarrhea			Whooping cough		
	Mortality	Absolute decreased value (%)	Relative decreased value (%)	Mortality	Absolute decreased value (%)	Relative decreased value (%)
1975	14.5			2.8		
1980	9.5	5.0	34.5	1.6	1.2	42.9
1985	3.7	5.8	61.1	0.9	0.7	43.8
1990	1.6	2.1	56.8	0.4	0.5	55.6
1995	0.7	0.9	56.3	0.2	0.2	50.0
2000	0.4	0.3	42.9	0.1	0.1	50.0

subject to the other things sequentially changing. The vertical and horizontal axes use the linear scale in the general line chart.

- (5) Box plot: It is useful to compare the average level and variation among different groups. It displays the minimum value, lower quartile (QL),

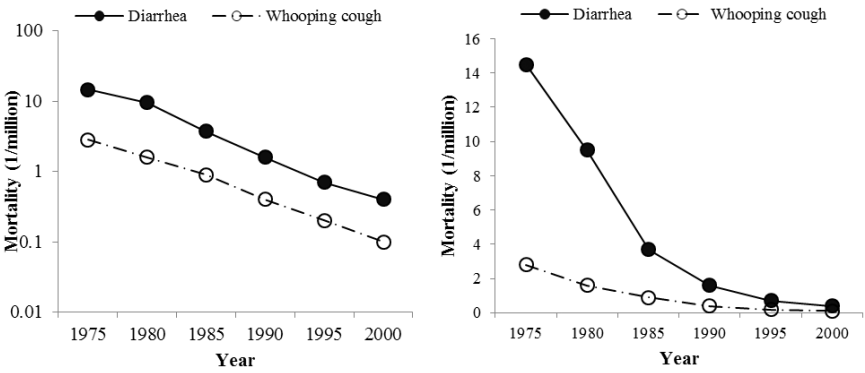


Fig. 1.15 The mortalities of diarrhea and whooping cough at one place from 1975 to 2000 (1/million).

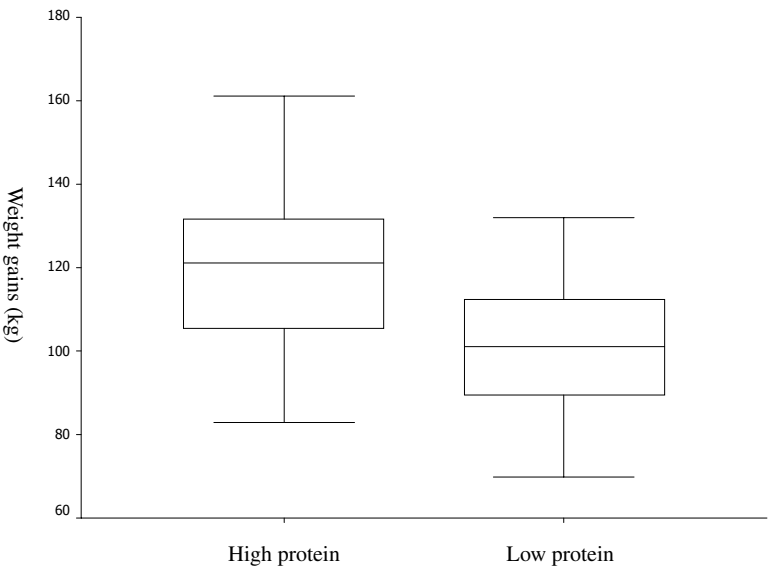


Fig. 1.16 The weight gains after using two kinds of feed with different protein content.

median (M), upper quartile (QU) and maximum value within each group. M indicates the average level; the range between minimum and maximum values, and the interquartile between QL and QU indicate the variation of the data.

Frequency	Stem & Leaf
2.00	5 . 3
9.00	5 . 4555
28.00	5 . 666666777777
21.00	5 . 888889999
39.00	6 . 00000000001111111
51.00	6 . 2222222222333333333333
82.00	6 . 4444444444444444555555555555555555555555
53.00	6 . 66666666677777777777777777777777
64.00	6 . 88888888888888888888889999999999
29.00	7 . 00000000111111
35.00	7 . 22222222223333333
28.00	7 . 4444444445555
21.00	7 . 666677777
6.00	7 . 889
2.00	8 . 0
7.00	Extremes (>=82)
(Stem width: 10.0; Each leaf: 2 case(s))	

Fig. 1.17 The weight distribution of 477 cesarean maternity patients.

- (6) **Stem-and-leaf plot:** The stem-and-leaf plot is similar to histogram. The stem-and-leaf display is drawn with two columns. The leaves are listed to the right, and contain all the last digit of the number; the stems are listed to the left, and contain all the other digits. It assists to quantitatively visualize the shape of a distribution. We can get each value from the stem-and-leaf plot. For example, in Fig. 1.17, the two values in the first group should be $5 \times 10 + 3 = 53$ (kg).

1.7 Computerized Experiments

Experiment 1.1 Frequency table, histogram and cumulative frequency plot 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:

- (1) Find out the minimum and maximum: Lines 01–05, read the data and find out the minimum and maximum.

(2) Design the subgroups: By calculating the range and deciding the number of subgroups, it is obtained as follows:

Subgroup	Mid-value	Subgroup	Mid-value
3.20–	3.35	4.70–	4.85
3.50–	3.65	5.00–	5.15
3.80–	3.95	5.60–	5.75
4.10–	4.25	5.90–6.20	6.05
4.40–	4.55		

(3) Organize data and list frequency table: Lines 08–21, each value is changed with the corresponding mid-value of its subgroup; lines 22–25 calculate description statistics such as mean, variance, standard deviation and variation coefficient (although the median and quartile could be given, they are just the mid-values in their sub-intervals instead

Program 1.1 Frequency table and histogram.

Line	Program	Line	Program
01	DATA RBC;	16	IF X<4.70 & X>=4.40 THEN Y=4.55;
02	INPUT X @@;	17	IF X<5.00 & X>=4.70 THEN Y=4.85;
03	CARDS;	18	IF X<5.30 & X>=5.00 THEN Y=5.15;
04	5.12 5.13	19	IF X<5.60 & X>=5.30 THEN Y=5.45;
05	20	IF X<5.90 & X>=5.60 THEN Y=5.75;
06	3.86 5.69	21	IF X>5.90 THEN Y=6.05;
07	;	22	PROC UNIVARIATE FREQ;
08	PROC MEANS MIN MAX;	23	VAR Y;
09	RUN;	24	RUN;
10	DATA FRBC;	25	PROC UNIVARIATE;
11	SET RBC;	26	CDF X / NORMAL; RUN;
12	IF X<3.50 THEN Y=3.35;	27	PROC GCHART;
13	IF X<3.80 & X>=3.50 THEN Y=3.65;	28	VBAR Y/TYPE=PERCENT;
14	IF X<4.10 & X>=3.80 THEN Y=3.95;	29	VBAR Y/TYPE=CPERCENT;
15	IF X<4.40 & X>=4.10 THEN Y=4.25;	30	RUN;

of the values obtained by interpolation introduced above) and the frequency table is performed.

- (4) Histogram and cumulative frequency plot: Lines 25 and 26 work out the cumulative frequency plot, and lines 27–30 work out the frequency distribution and histogram.

Experiment 1.2 Clustered bar chart The detailed steps in the software for clustered bar chart are as follows (Program 1.2):

- (1) Data input: Lines 01–09, input the data as the following table: The resistance rates of five types of staphylococcus for two kinds of antibiotics

Types of staphylococcus	Penicillin (%)	Cephalothin V (%)
Golden	81.5	37.7
Epidermidis	81.3	35.1
Saprophytes	80.5	27.5
Haemolyticus	94.5	66.0
Simulans	92.3	62.1

- (2) Define the chart's structure: Lines 10–13 define the *X*-axis and *Y*-axis. *X*-axis is labeled as "Category of staphylococcus"; *Y*-axis is labeled as "RATE", ranged from 0% to 100%, 10% as the length of interval.
- (3) Label the variables: Lines 13–15, Label a1–a5 and a, b as the corresponding five categories of staphylococcus and two types of antibiotics, respectively.
- (4) Plot the chart: Line 16 is the plot PROC step; lines 17 and 18 indicate that the clustered bar chart should be plot based on the categories of staphylococcal bacteria and two types of antibiotics; line 19 defines the labels in step (3); lines 20–22 define the shape and the color in the chart.

Experiment 1.3 Pie chart The SAS codes for the pie chart is in Program 1.3. Firstly, line 01 sets the background of the chart, lines 02–12 input the data. Lines 13–15 make the frequency table of the dataset named JOB, and output the frequency table to the dataset named JOBPCT.

Program 1.2 Clustered bar chart.

Line	Program	Line	Program
01	DATA ANTIBIO;	13	PROC FORMAT;
02	INPUT BA\$ ANTI\$ RATE@@;	14	VALUE \$ss a1="G" a2="E"
03	CARDS;		a3="S" a4="H" a5="SM";
04	a1 a 81.5 a1 b 37.7	15	VALUE\$qq a="penicillin"
05	a2 a 81.3 a2 b 35.1		b="cephalothin V";
06	a3 a 80.5 a3 b 27.5	16	PROC GCHART;
07	a4 a 94.5 a4 b 66.0	17	WHERE ANTI in ("a", "b");
08	a5 a 92.3 a5 b 62.1	18	VBAR ANTI/GROUP=BA
09	;		SUMVAR=RATE
10	GOPTIONS RESET=ALL;		ATTERNID=MIDPOINT;
11	AXIS1 LABEL=('staphylococcus')	19	FORMAT BA \$SS. ANTI \$QQ.;
	VALUE=('H SM G E S')	20	PATTERN1 V=L5 C=GRAY;
12	AXIS2 LABEL=('RATE')	21	PATTERN2 V=X5 C=GRAY;
	VALUE=(0 TO 100 BY 10);	22	RUN;

Program 1.3 Pie chart.

Line	Program	Line	Program
01	GOPTIONS RESET=ALL	19	PATTERN1 V=P3N0 C=GRAY;
	CBACK=WHITE BORDER	20	PATTERN2 V=E C=GRAY;
	HTITLE=12pt HTEXT 10pt;	21	PATTERN3 V=P3N45 C=GRAY;
02	DATA JOB;	22	PATTERN4 V=P3X45 C=GRAY;
03	LENGTH WORK \$8 ;	23	PATTERN5 V=P3N90 C=GRAY;
04	INPUT ID WORK;	24	PATTERN6 V=S C=GRAY;
05	DATALINES;	25	LEGEND1 LABEL=NONE
06	1 workers		POSITION=(RIGHT MIDDLE)
07	2 others		OFFSET=(,4) ACROSS=1
08	3 intellectual		VALUE=(COLOR=BLACK)
09	4 farmers		SHAPE=BAR(4,1.5);
10	5 services	26	PROC GCHART DATA=JOBPCT;
11	...	27	PIE WORK/SUMVAR=PERCENT
12	;		SLICE=INSIDE
13	PROC FREQ DATA=JOB;		PERCENT=INSIDE
14	TABLES WORK/OUT=JOBPCT;		LEGEND=LEGEND1
15	RUN;		MIDPOINTS='worker' 'others'
16	ODS RTF;		'farmers' 'managers' 'intellectual'
17	ODS GRAPHICS ON;		'services' NOHEADING;
18	TITLE 'JOB PERCENTAGE';	28	RUN;
		29	ODS GRAPHICS OFF;
		30	ODS RTF CLOSE;

Lines 16 and 17 define the output format of the pie chart as word file. Then line 8 defines the title, lines 19–24 define the patterns and the colors for each slice: V define THE dark or light and the patterns for each slice, and C defines the colors for each slice. Line 25 defines the legends: LEBEL defines the names of the legends, POSITION defines the positions of the legends, OFFSET defines the distance between the legends and the edge of the chart, ACROSS defines the amount of the legends (only one here), COLOR in VALUE defines the text color of the legends, and SHAPE defines the size of the legends.

Finally, lines 26–28 plot the pie chart based on the new dataset JOBPCT. MIDPOINTS define the slices are anti-clockwise ordered (PATTERN defines the slices' pattern according to the anti-clockwise order). Lines 29–30 complete the output.

Experiment 1.4 Box plot The SAS codes for the box plot is in Program 1.4. Line 01 sets the background for the plot; lines 02–07 input the data; lines 08–09 define the output format of the plot as word file, line 10 defines the title of the box plot. Line 11 defines the patterns of the plot: INTERPOL=BOXT5 defines the 95% percentile as the upper whisker and 5% percentile as the lower whisker. WIDTH defines the width of the box.

Program 1.4 Box plot.

Line	Program	Line	Program
01	GOPTIONS RESET=ALL CBACK=WHITE BORDER HTITLE=12pt HTEXT=10pt;	11	SYMBOL INTERPOL=BOXT5
02	DATA PROTEIN;	12	WIDTTH=10; AXIS1 LABEL=NONE VALUE=(T=1 'high protein' T=2 'low protein')
03	INPUT GROUP \$ WEIGHT @@;		OFFSET= (5,5) LENGTH=50;
04	DATALINES;	13	AXIS2 LABEL= ('gain weight(g)') MINOR=NONE
05	A 134 A 146 A 104 A 119 A 124. . .		ORDER= (60 TO 180 BY20);
06	B 70 B 118 B 101 B 85 B 107. . .;	14	PROC GPLOT DATA=PROTEIN;
07	RUN;	15	PLOT WEIGHT*GROUP/HAXIS=AXIS1 VAXIS=AXIS2;
08	ODS RTF;	16	RUN;
09	ODS GRAPHICS ON;	17	ODS GRAPHICS OFF;
10	TITLE1 'COMPARISON: WEIGHT BY GROUP';	18	ODS RTF CLOSE;

Program 1.5 Program for direct and indirect approaches.

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

Lines 12 and 13 define the vertical and horizontal axis; lines 14–16 plot the box plot, finally lines 7–18 complete the output.

Experiment 1.5 Calculation of standardized mortality rate with direct and indirect approaches Program 1.2 is the SAS program for reference. The first 20 lines are for the direct approach where lines 4 and 5 calculate the age specific mortality rates, lines 7 and 8 calculate the age specific numbers of deaths, lines 10–17 list the data, and lines 18 and 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 and 25

calculate the age specific expected numbers of deaths; lines 27–34 list the data; lines 35–37 calculate the 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.

1.8 Practice and Experiments

1. True or false: Which of the following statements are correct?
 - (1) “The red blood cells in occult blood examination” is a continuous variable.
 - (2) Red blood cell count is a discrete variable.
 - (3) The arithmetic mean is always greater than the median.
 - (4) The mean of large sample is always 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 cumulative frequency curve is a stepwise curve where the values are sparse.
 - (8) The distribution of the days of hospitalization for certain disease is 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.
 - (9) The dimension of variation of coefficient is the same as that of the original variable.
 - (10) If the sample mean is greater, then the standard deviation must be greater.
 - (11) 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. Calculate the arithmetic mean, geometric mean and median; which answer better reflects the average level? Again calculate the range, $Q_3 - Q_1$ and standard deviation; which answer better reflects the variation?
4. The daily fat intake (g) of 100 randomly selected adults was surveyed with the data as 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

- Work out a frequency table, a histogram, box and whiskers plot, and stem and leaf plot; 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 blood 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:

Cumulated incidence rate

$$= \frac{\text{Number of new patients during the same period}}{\text{Total number of persons followed during certain period}}$$

Person-year incidence rate

$$= \frac{\text{Number of new patients during the same period}}{\text{Total person-years of exposure to the risk during certain period}}$$

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 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	0.0859	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

Compare the risk of liver cancer 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. Compare the risk of liver cancer 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 cities *A* and *B* as a standard population;
 - (4) Compare and discuss the results.
9. What are the frequently used graphs in statistics? What are the different situations for the use of different types of graphs?

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

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

$$(3) \sum_{i=1}^n (a + y_i) = na + \sum_{i=1}^n y_i; \quad (4) \sum_{i=1}^n \left(\frac{y_i}{n} a \right) = a \sum_{i=1}^n \frac{y_i}{n};$$

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

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

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

(1st edn. Jiqian Fang; 2nd edn. Chun Hao, Jiqian Fang)