# Chapter 1
# Warming Up: Descriptive Statistics and Essential Probability Models

This chapter portrays how to make sense of gathered data before performing the formal statistical inference. The covered topics are types of data, how to visualize data, how to summarize data into few descriptive statistics (i.e., condensed numerical indices), and introduction to some useful probability models.

## 1.1 Types of Data

Typical types of data arising from clinical studies mostly fall into one of the following categories.

Nominal categorical data contain qualitative information and appear to discrete values that are codified into numbers or characters (e.g., 1=case with a disease diagnosis, 0 = control; M = male, F = female).

Ordinal categorical data are *semi*-quantitative and discrete, and the numeric coding scheme is to order the values such as 1 = mild, 2 = moderate, and 3 = severe. Note that the value of 3 (severe) does not necessarily be three times more severe than 1 (mild).

Count (number of events) data are quantitative and discrete (i.e., 0, 1, 2 …).

Interval scale data are quantitative and continuous. There is no absolute 0 and the reference value is arbitrary. Particular examples of such data are temperature values in °C and °F.

Ratio scale data are quantitative and continuous, and there is the absolute 0. Particular examples of such data are body weight and height.

In most cases the types of data usually fall into the above classification scheme shown in Table 1.1 in that the types of data can be classified into either quantitative or qualitative, and discrete or continuous. Nonetheless, some definition of the data type may not be clear and among which the similarity and dissimilarity between the ratio scale and interval scale may be such ones that need further clarification.

H. Lee, *Foundations of Applied Statistical Methods*, DOI 10.1007/978-3-319-02402-8_1,

© Springer International Publishing Switzerland 2014

**Table 1.1** Classifications of data types

| | Qualitative | Quantitative |
|---|---|---|
| Discrete | Nominal categorical (e.g., M=male, F=female) | Ordinal categorical (e.g., 1=mild, 2=moderate, 3=severe) |
| | | Count (e.g., number of incidences 0, 1, 2, 3, …) |
| Continuous | N/A | Interval scale (e.g., temperature) |
| | | Ratio scale (e.g., weight) |

Ratio scale: Two distinct values of the ratio scale are ratio-able. For example, the ratio of two distinct values of a ratio scale $x$, $x_1/x_2 = 2$ for $x_1 = 200$ and $x_2 = 100$, can be interpreted as "twice as large." Blood cholesterol level, measured as the total volume of cholesterol molecule in a certain unit, is such an example in that if person A's cholesterol level to person B's cholesterol level ratio is 2, then we will be able to say that person A's cholesterol level is doubly higher than that of person B. Other such examples are lung volume, age, and disease duration.

Interval scale: If two distinct values of quantitative data were not ratio-able, then such data are interval scale data. Temperature is a good example in that there are three temperature systems, i.e., Fahrenheit, Celsius, and Kelvin. Kelvin system even has its absolute 0 (there is no negative temperature in Kelvin system). For example, 200 °F is not a temperature that is twice higher than 100 °F. We can only say that 200° is higher by 100° (i.e., the displacement between 200° and 100° is 100° in the Fahrenheit measurement scale).

## 1.2 Description of Data Pattern

### *1.2.1 Distribution*

A distribution is a complete description of how large the occurring chance (i.e., probability) of a unique datum or certain range of data is. The following two explanations will help you grasp the concept. If you keep on rolling a die, you expect to observe 1, 2, 3, 4, 5, or 6 equally likely, i.e., a probability for each unique outcome value is 1/6. We say "a probability of 1/6 is distributed to the value of 1, 1/6 is distributed to 2, 1/6 to 3, 1/6 to 4, 1/6 to 5, and 1/6 to 6, respectively." Another example is that if you keep on rolling a die many times, and each time you say "a success" if the observed outcome is 5 or 6 and say "a failure" otherwise, then your expected chance to observe a success is 1/3 and that of a failure is 2/3. We say "a probability of 1/3 is distributed to the success and 2/3 is distributed to the failure". In real life, there are many distributions that cannot be verbalized as simply as these two examples, which require descriptions using sophisticated mathematical expressions.

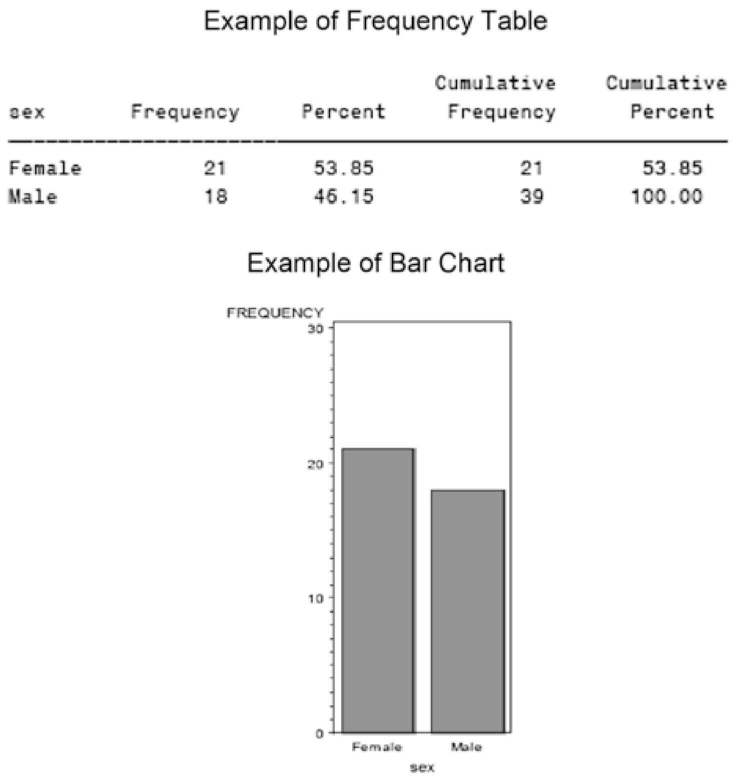


**Fig. 1.1** Frequency table and bar chart for describing nominal categorical data

Example of Frequency Table

```
                                           Cumulative    Cumulative
sex        Frequency     Percent           Frequency       Percent
---------------------------------------------------------------------
Female            21       53.85                  21         53.85
Male              18       46.15                  39        100.00
```

Example of Bar Chart

Let's discuss how to describe the distributions arising from various types of data. One way to describe a set of collected data is to make description about the distribution of relative frequency for the observed individual values (e.g., what values are how much common and what values are how much less common). Graphs, simple tables, or a few summary numbers are commonly used.

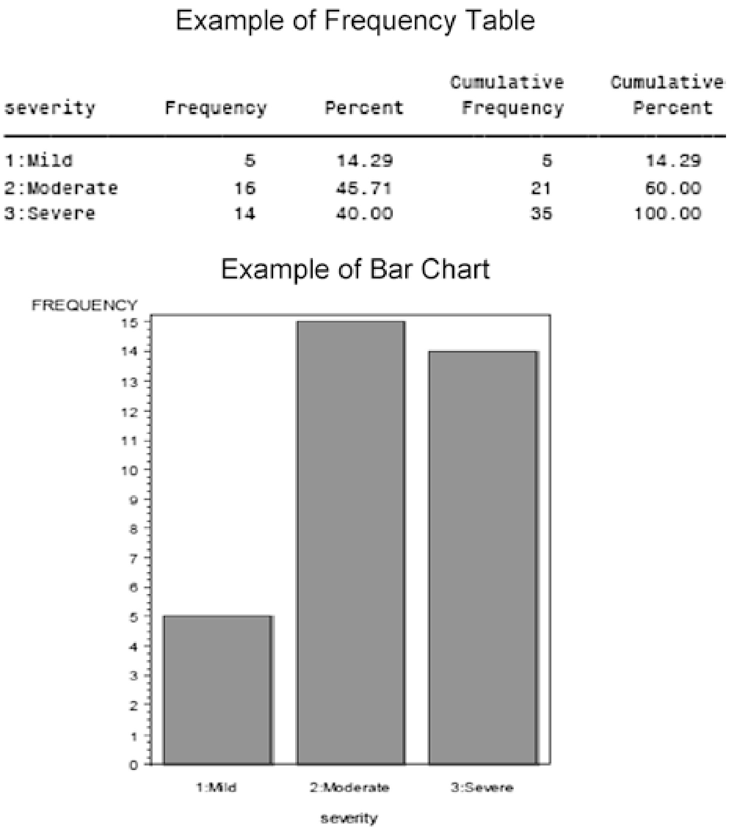
## *1.2.2 Description of Categorical Data Distribution*

A simple tabulation, *aka* frequency table, is to list the observed count (and proportion in percentage value) for each category. A bar chart (see Figs. 1.1 and 1.2) is a good visual description of where the horizontal axis defines the categories of the outcome and the vertical axis shows the frequency of each observed category. The size of each bar in the Figures is the actual count. It is also common to present the relative frequency (i.e., proportion of each category in percentage value).

## *1.2.3 Description of Continuous Data Distribution*

Figure 1.3 is a listing of white blood cell (WBC) counts of 31 patients diagnosed with a certain illness listed by the patient identification number. Does this listing itself tell us the group characteristics such as the average and the variability among patients?

**Fig. 1.2** Frequency table and bar chart for describing ordinal data



Example of Frequency Table

```
                                             Cumulative    Cumulative
severity        Frequency     Percent         Frequency       Percent
----------------------------------------------------------------------
1:Mild                 5        14.29                5         14.29
2:Moderate            16        45.71               21         60.00
3:Severe              14        40.00               35        100.00
```

Example of Bar Chart

**Fig. 1.3** List of WBC raw data of 31 subjects

```
 ID      WBC

  1     5200
  2     3100
  3     3000
  4     3700
  5     4000
  6     3700
  7     4100
  8     8100
  9     3500
 10     3300
 11     3400
 12     9300
 13     3800
 14     2600
 15     4800
 16     5800
 17     4300
 18     6100
 19     6100
 20     1800
 21     7500
 22     4100
 23    11200
 24     2800
 25     3900
 26     3400
 27     8900
 28     5900
 29     4500
 30     3800
 31     6500
```

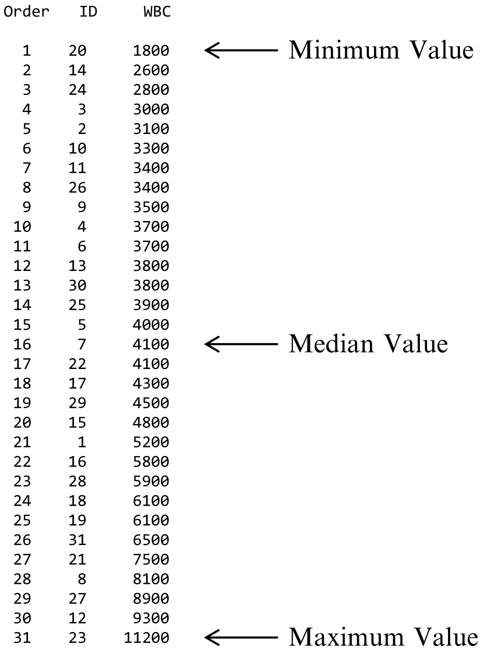


**Fig. 1.4** List of 31 individual WBC values in ascending order

| Order | ID | WBC |
|---|---|---|
| 1 | 20 | 1800 |
| 2 | 14 | 2600 |
| 3 | 24 | 2800 |
| 4 | 3 | 3000 |
| 5 | 2 | 3100 |
| 6 | 10 | 3300 |
| 7 | 11 | 3400 |
| 8 | 26 | 3400 |
| 9 | 9 | 3500 |
| 10 | 4 | 3700 |
| 11 | 6 | 3700 |
| 12 | 13 | 3800 |
| 13 | 30 | 3800 |
| 14 | 25 | 3900 |
| 15 | 5 | 4000 |
| 16 | 7 | 4100 |
| 17 | 22 | 4100 |
| 18 | 17 | 4300 |
| 19 | 29 | 4500 |
| 20 | 15 | 4800 |
| 21 | 1 | 5200 |
| 22 | 16 | 5800 |
| 23 | 28 | 5900 |
| 24 | 18 | 6100 |
| 25 | 19 | 6100 |
| 26 | 31 | 6500 |
| 27 | 21 | 7500 |
| 28 | 8 | 8100 |
| 29 | 27 | 8900 |
| 30 | 12 | 9300 |
| 31 | 23 | 11200 |

Minimum Value
Median Value
Maximum Value

How can we describe the distribution of these data, i.e., how much of the occurring chance is distributed to WBC=5,200, how much to WBC=3,100 …, and etc.? Such a description may be very cumbersome. As depicted in Fig. 1.4, the listed full data in ascending order can be a primitive way to describe the distribution, but it does not still describe the distribution. An option is to visualize the relative frequencies for grouped intervals of the observed data. Such a presentation is called histogram. To create a histogram, one will first need to create equally spaced WBC categories and count how many observations fall into each category. Then the bar graph can be drawn where each bar size indicates the relative frequency of that particular WBC interval category. This may be a daunting task. Rather than covering the techniques to create the histogram, next section introduces an alternative option.

### *1.2.4 Stem-and-Leaf*

The Stem-and-Leaf plot requires much less work than creating the conventional histogram while providing the same information as what the histogram does. This is a quick and easy option to sketch a continuous data distribution.

Let's use a small data set for illustration, and then revisit our WBC data example for more discussion (Fig. 1.10) after we become familiar to this method. The following nine data points: 12, 32, 22, 28, 26, 45, 32, 21, and 85, are ages (ratio scale) of a small group. Figures 1.5, 1.6, 1.7, 1.8, and 1.9 demonstrate how to create the Stem-and-Leaf plot of these data.

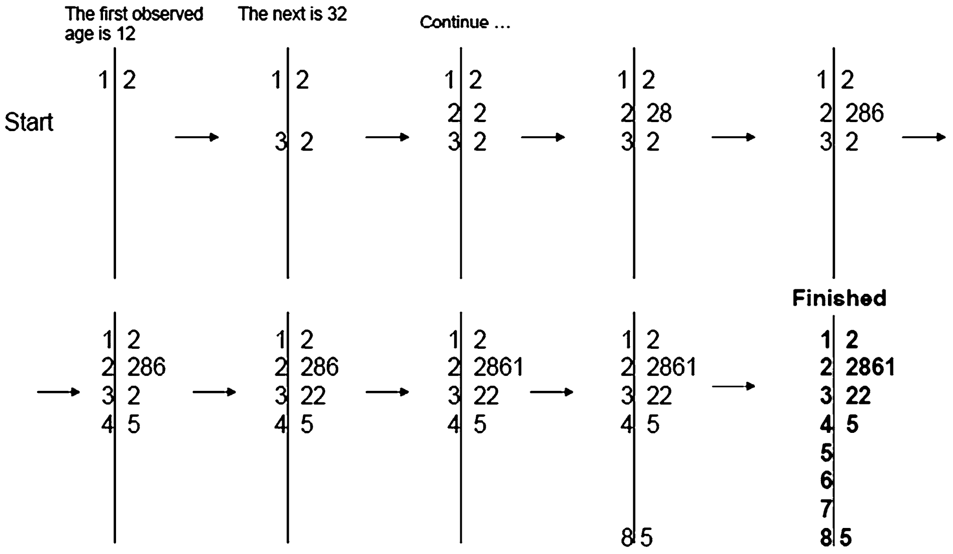


**Fig. 1.5** Step-by-step illustration of creating a Stem-and-Leaf plot

**Fig. 1.6** Illustration of creating a Stem-and-Leaf plot

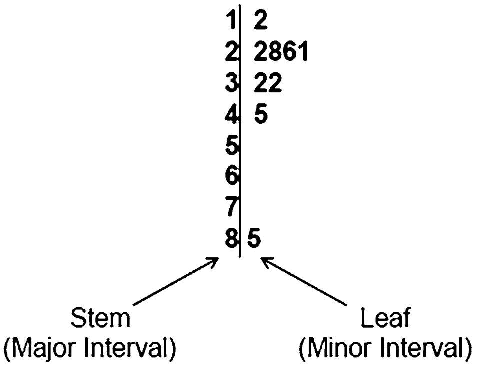
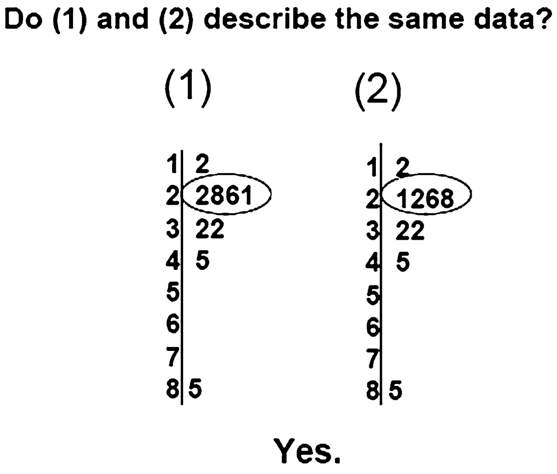
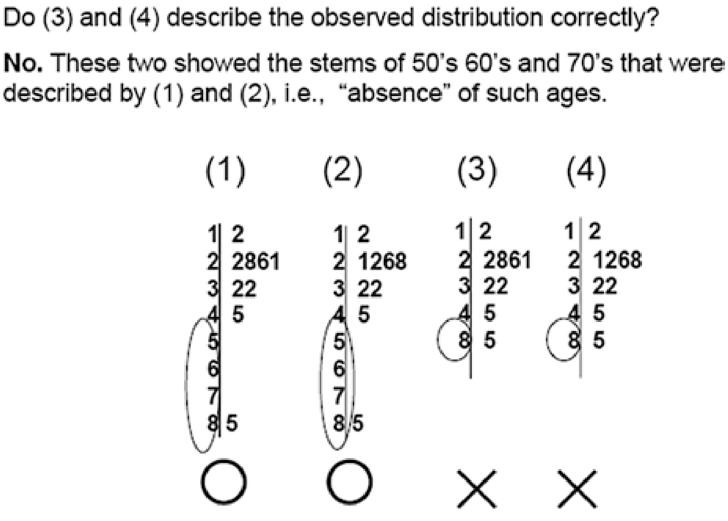


**Fig. 1.7** Two Stem-and-Leaf plots describing the same data

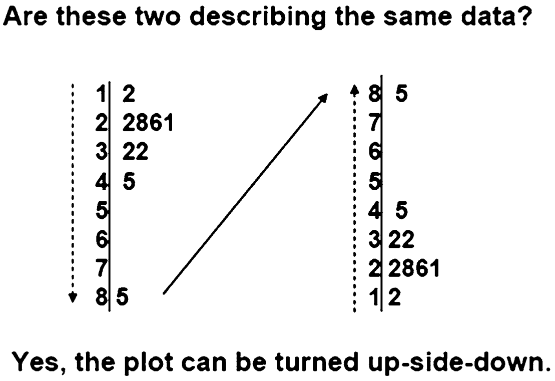




Do (3) and (4) describe the observed distribution correctly?

**No.** These two showed the stems of 50's 60's and 70's that were described by (1) and (2), i.e., "absence" of such ages.

**Fig. 1.8** Common mistakes in Stem-and-Leaf



**Are these two describing the same data?**

**Yes, the plot can be turned up-side-down.**

**Fig. 1.9** Two Stem-and-Leaf plots describing the same distribution by ascending and descending orders

The main idea of this technique is a quick sketch of the distribution of an observed data set without computational burden. Let's just take each datum in the order that it is recorded (i.e., the data are not preprocessed by other techniques such as sorting by ascending/descending order) and plot one value at a time (see Fig. 1.5). Note that the oldest observed age is 85 years which is much greater than the next oldest age 45 years, and the unobserved stem interval values (i.e., 50s, 60s, and 70s) are placed. The determination of the number of equally spaced major intervals (i.e., number of stems) can be subjective and data range-dependent.

As presented in Fig. 1.10, the distribution of our WBC data set is described by the Stem-and-Leaf plot. Noted observations are: most values lie between 3,000 and 4,000 (i.e., mode); the contour of the frequency distribution is skewed to the right and the mean value did not describe the central location well; and the smallest and the largest observations were 1,800 and 11,200, respectively.

```
Stem-Leaf*                Frequency**

    11-2                       1
    10-
    9-3                        1
    8-19                       2
    7-5                        1
    6-115                      3
    5-289                      3
    4-011358                   6
    3-01344577889             11
    2-68                       2
    1-8                        1

*Multiply Stem-Leaf by 1000 Multiply Stem-Leaf by 1000
```

** Frequency counts annotation is not a part of the Stem-and-Leaf and unnecessary but presented to aid the reading.

**Fig. 1.10** Presentation of WBC data of 31 subjects using Stem-and-Leaf

## 1.3 Descriptive Statistics

In addition to the visual description such as Stem-and-Leaf plot, further description of the distribution by means of a few statistical metrics is useful. Such metrics are called descriptive statistics which indicate where most of the data values are concentrated and how much the occurring chance distribution is scattered around that concentrated location.

### *1.3.1 Statistic*

A **statistic** is a function of data, wherein a function usually appears as a mathematical expression that takes the observed data and reduces to a single summary metric, e.g., mean = sum over all data divided by the number of sample size. Note that the word mathematical expression is interchangeable with formula. As the word formula is usually referred in a plug-and-play setting, this monograph names it mathematical expression, and the least amount of the expression is introduced only when necessary.

### *1.3.2 Central Tendency Descriptive Statistics for Quantitative Outcomes*

In practice, there are two kinds of descriptive statistics used for quantitative outcomes of which the one kind is the metric indices for characterizing the central tendency and the second is for the dispersion. The mean (i.e., sum of all observations divided by the sample size), the median (i.e., the midpoint value), and the mode (i.e., the most frequent value) are the central tendency descriptive statistics.

### 1.3.3 Dispersion Descriptive Statistics for Quantitative Outcomes

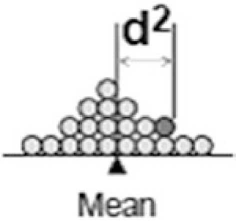
The range (i.e., maximum value–minimum value) and interquartile range (i.e., 75th–25th percentile) are very simple to generate by which the dispersion of a data set is described. Other commonly used dispersion descriptive statistics are variance, standard deviation, and coefficient of variation, and these describe the dispersion of data (particularly when the data are symmetrically scattered around the mean), and the variance and standard deviation are important statistics that play a pivotal role in the formal statistical inferences which will be discussed in Chap. 2.

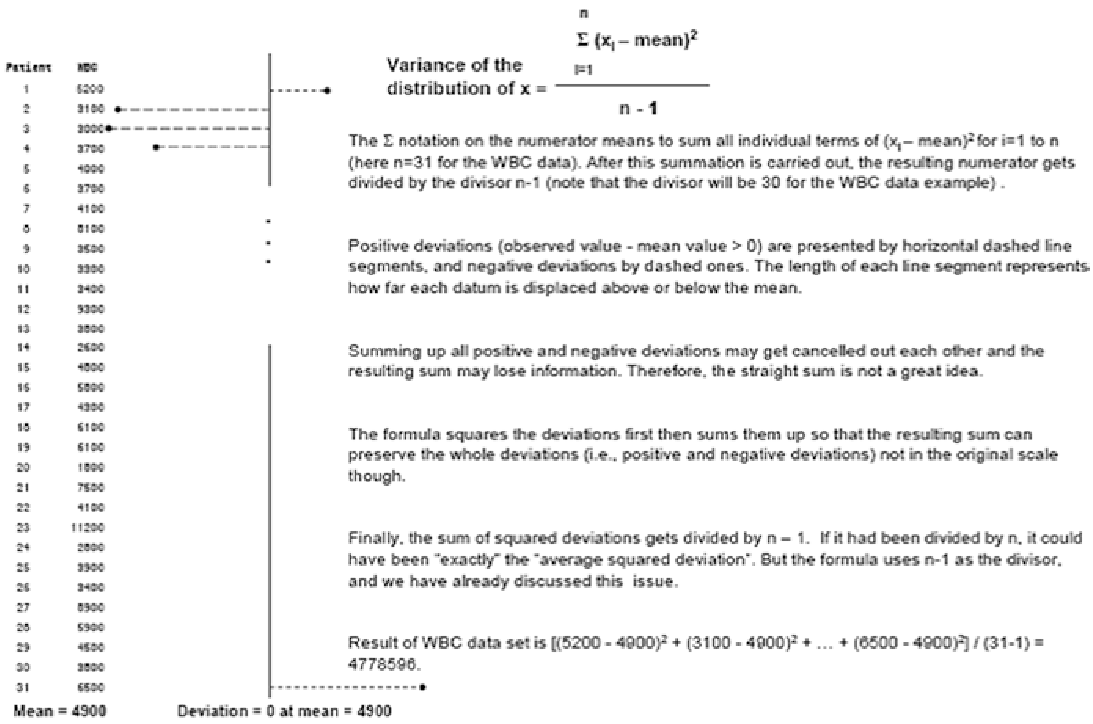
### 1.3.4 Variance

The variance of a distribution, denoted by $\sigma^2$, can be conceptualized an average squared deviation (explained in detail below) of the data values from their mean. The more dispersed the data are, the more the variance increases. It is common that standard textbooks present the *definitional* and *computational* mathematical expressions. Until the modern computer was not widely available, finding a shortcut for manual calculations and choosing a right tool for a quick and easy calculation had been a major issue of statistical education and practice. Today's data analysis utilizing computer software and knowledge about the shortcut for manual calculations is not important. Nonetheless, understanding the genesis of definitional expression, at least, is important. The following is the demonstration of the definitional expression of the variance.

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_\mathrm{i} - \bar{x})^2}{n-1},$$

where $x_\mathrm{i}$'s, for $i=1, 2, \ldots n$ (i.e., the sample size) are the individual data values, $\bar{x}$ is their mean. The $\sum_{i=1}^{n}$ notation on the numerator is to sum over all individual terms, $(x_i - \bar{x})^2$, for $i$ = 1 to $n$ (e.g., $n$ = 31 for the WBC data). The term $(x_i - \bar{x})^2$ for $i$ is the squared deviation of an individual data value from its mean and is depicted by $\mathbf{d}^2$ in the following visual demonstration.





| Patient | WBC |
|---|---|
| 1 | 5200 |
| 2 | 3100 |
| 3 | 3000 |
| 4 | 3700 |
| 5 | 4000 |
| 6 | 3700 |
| 7 | 4100 |
| 8 | 8100 |
| 9 | 3500 |
| 10 | 3300 |
| 11 | 3400 |
| 12 | 9300 |
| 13 | 3800 |
| 14 | 2600 |
| 15 | 4800 |
| 16 | 5800 |
| 17 | 4300 |
| 18 | 6100 |
| 19 | 6100 |
| 20 | 1800 |
| 21 | 7500 |
| 22 | 4100 |
| 23 | 11200 |
| 24 | 2800 |
| 25 | 3900 |
| 26 | 3400 |
| 27 | 8900 |
| 28 | 5900 |
| 29 | 4600 |
| 30 | 3800 |
| 31 | 6500 |

Mean = 4900

Deviation = 0 at mean = 4900

Variance of the distribution of x = $\dfrac{\sum_{i=1}^{n} (x_i - \text{mean})^2}{n - 1}$

The Σ notation on the numerator means to sum all individual terms of $(x_i - \text{mean})^2$ for i=1 to n (here n=31 for the WBC data). After this summation is carried out, the resulting numerator gets divided by the divisor n-1 (note that the divisor will be 30 for the WBC data example) .

Positive deviations (observed value - mean value > 0) are presented by horizontal dashed line segments, and negative deviations by dashed ones. The length of each line segment represents how far each datum is displaced above or below the mean.

Summing up all positive and negative deviations may get cancelled out each other and the resulting sum may lose information. Therefore, the straight sum is not a great idea.

The formula squares the deviations first then sums them up so that the resulting sum can preserve the whole deviations (i.e., positive and negative deviations) not in the original scale though.

Finally, the sum of squared deviations gets divided by n – 1. If it had been divided by n, it could have been "exactly" the "average squared deviation". But the formula uses n-1 as the divisor, and we have already discussed this issue.

Result of WBC data set is $[(5200 - 4900)^2 + (3100 - 4900)^2 + \ldots + (6500 - 4900)^2] / (31\text{-}1) =$ 4778596.

**Fig. 1.11** Definitional formula of variance

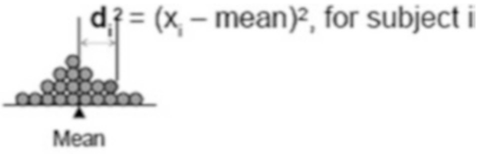
After this summation is carried out, the resulting numerator is then divided by the divisor *n – 1* (note that the divisor will be 30 for the WBC data example).

As depicted in Fig. 1.11, positive deviations (i.e., $x_i - \overline{x} > 0$) are presented by horizontal dashed line segments, and negative deviations (i.e., $x_i - \overline{x} < 0$) by dashed ones. The length of each line segment represents how far each datum is displaced above or below the mean. How do we cumulate and preserve the deviations of the entire group? If straight summation is considered, the positive and negative individual deviations may get cancelled out each other and the resulting sum may not retain the information. Thus the straight summation is not a great idea. The individual deviations are squared first then summed up so that the resulting sum can retain the information (i.e., positive and negative deviations) although the retained quantity is not in the original scale. Then, the sum of squared deviations is divided by *n – 1*. If it had been divided by *n*, it could have been literally the average squared deviation. Instead, the used divisor is *n-1*. Normally an average is obtained by dividing the sum of all values by the sample size *n*. However, when computing the variance using sample data, we divide by *n-1*, not by *n*. The idea behind is the following. If the numerator (i.e., sum of squared deviations from the mean) is divided by the sample size, *n*, then such a calculation will slightly downsize the true standard deviation. The reason is that when the deviation of each individual data point from the mean was obtained, the mean is usually not externally given to us but is generated within the given data set and thus the actually observed deviations could become slightly smaller than what it should be (i.e., referencing to an internally obtained mean value). So, in order to make an appropriate adjustment for the final averaging step, we divide it by *n -1*. You may be curious why it has to be 1 less than the sample size, not 2 less than, or something else. We can at least show that 2 less cannot handle when the sample size is 2, and 3 less cannot handle the sample size of 3. Unlike other choices, *n-1* (i.e., 1 less than the sample size) can handle any sample size because the smallest sample size that will have a variance is 2 (obviously there is no variance for a single observation)? There is a formal proof that the divisor of *n-1* is the best for any sample size but it is not necessary to cover it in full detail within this introductory course setting.

The computed variance of the WBC data set is [(5200 - 4900)$^2$ + (3100 - 4900)$^2$ + … + (6500 - 4900)$^2$]/(31-1) = 4778596. Note that variance's unit is not the same as the raw data unit (because of the squaring the summed deviations).

### 1.3.5 Standard Deviation

The standard deviation of a distribution, denoted by σ, is the square root of variance (i.e., $\sqrt{variance}$), and the scale of the standard deviation is the same as that of the raw data. The greater the data are dispersed the standard deviation increases. If the dispersed data form a particular shape (e.g., bell curve), then one standard deviation unit symmetrically around (i.e., above and below) the mean will cover about middle two-thirds of the data range value (see standard normal distribution in Sect. 1.4.3).



$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} d_i^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \text{mean})^2}{n-1}}$$

### 1.3.6 Property of Standard Deviation After Data Transformations

The observed data often require transformations for analysis purposes. One example is to shift the whole data set to a new reference point by simply adding a positive constant to or subtracting it from the raw data values. Such a simple transformation does not alter the distances between the individual data values thus the standard deviation remains unchanged (Fig. 1.12).

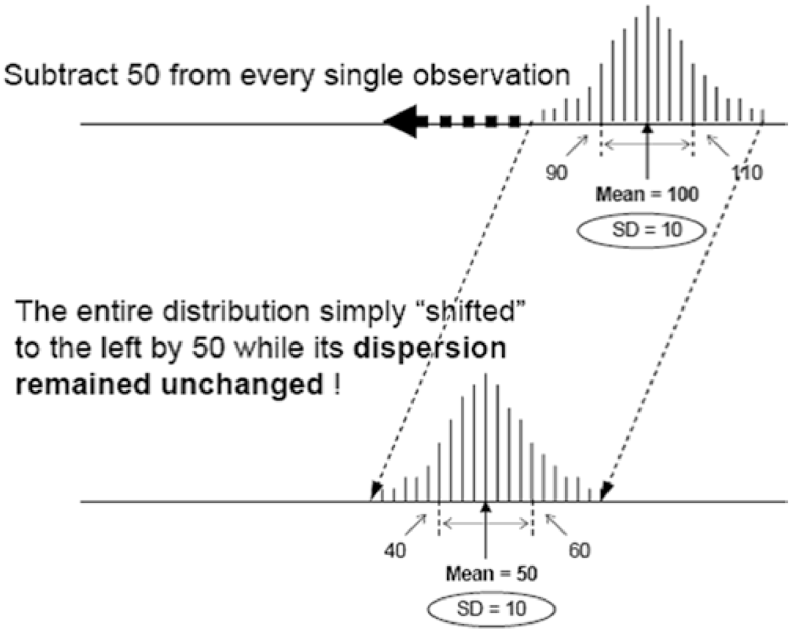

Subtract 50 from every single observation
90
110
Mean = 100
SD = 10
The entire distribution simply "shifted" to the left by 50 while its dispersion remained unchanged !
40
60
Mean = 50
SD = 10

**Fig. 1.12** Shifted data without changing dispersion

Another example is to change scale of the data without- or with changing the reference point. In general, if data *x* (a collection of $x_1$, $x_2$, …, $x_n$) of which the mean = μ and standard deviation = $\sigma_x$ is transformed to $y = a{\cdot}x+b$, where *a* is the scaling constant and *b* is the reference point, then the mean of *y* remains the same of $y = a{\cdot}(mean\ of\ x) + b = a{\cdot}\mu + b$ and the standard deviation *y*, $\sigma_y = a{\cdot}\sigma_x$. Note that adding a constant does not alter the original standard deviation, and only the scaling factor does.

The following example is to demonstrate how the means and standard deviations are changed after transformation. The first column lists a set of body temperature of eight individuals recorded in °C, the second column lists their deviations from the normal body temperature 36.5 °C (i.e., $d = C - 36.5$), and the third column lists their values in °F (i.e., $F = 1.8C + 32$). The mean of the deviations from the normal temperature is 0.33 (i.e., 0.33° higher than the normal temperature on average), which can be reproduced by the simple calculation of the difference between the two mean values 36.83 and 36.5 without having to recalculate the transformed individual data. The standard deviation remains the same because this transformation was just a shifting of the distribution to the reference point 32. The mean of the transformed values to °F scale is 98.29, which can be obtained by the simple calculation of 1.8 times the mean of 36.83 then add 32 without having to recalculate using the transformed individual observations. This transformation involves not only the distribution shifting but also the rescaling where the rescaling was to multiply the original observations by 1.8 prior to shifting the entire distribution to the reference point of 32. The standard deviation of the data transformed to °F scale is 1.12, which can be directly obtained by multiplying 1.8 to the standard deviation of the raw data in °C scale, i.e., 1.12 = 0.62 ×1.8 (Fig. 1.13).

| | Body Temperature °C (Raw Data) | Body Temperature Deviation from 36.5 °C Reference Point (Transformation: d = C - 36.5) | Body Temperature °F (Transformation: F = 1.8C + 32) |
|---|---|---|---|
| | 36.40 | -0.10 | 97.52 |
| | 36.50 | 0.00 | 97.70 |
| | 36.50 | 0.00 | 97.70 |
| | 36.50 | 0.00 | 97.70 |
| | 36.60 | 0.10 | 97.88 |
| | 37.20 | 0.70 | 98.96 |
| | 38.10 | 1.60 | 100.58 |
| Stem and Leaf | 38.(0~4) \| 1 | 1.(5~9) \| 6 | |
| | 37.(5~9) \| 2 | 1.(0~4) \| | 100. \| 6 |
| | 37.(0~4) \| | 0.(5~9) \| 7 | 99. \| 0* |
| | 36.(5~9) \| 5556 | 0.(0~4) \| 0001 | 98. \| |
| | 36.(0~4) \| 4 | -0.(0~4) \| 1 | 97. \| 0005 |
| | | | * 98.96 was rounded to 99.0 |
| Mean | 36.83 | 0.33 (subtract 36.5 from the original mean) | 98.29 |
| Std. Dev. | 0.62 | 0.62 (recalculation is unnecessary) | 1.12 (0.62 was multiplied by .8) |

**Fig. 1.13** Scale invariant and scale variant transformations

Mean: 4910
Median: 4100
Mode: 3500~3999

| Stem-Leaf* | Frequency** |
|---|---|
| 11-2 | 1 |
| 10- | |
| 9-3 | 1 |
| 8-19 | 2 |
| 7-5 | 1 |
| 6-115 | 3 |
| 5-289 | 3 |
| 4-011358 | 6 |
| 3-01344577889 | 11 |
| 2-68 | 2 |
| 1-8 | 1 |

*Multiply Stem-Leaf by 1000 Multiply Stem-Leaf by 1000

**Fig. 1.14** Asymmetrical distribution depicted by a Stem-and-leaf plot

## *1.3.7 Other Descriptive Statistics for Dispersion*

Figure 1.14 illustrates the asymmetrical distribution of the WBC that was illustrated in Fig. 1.10. The mean, median, and mode are not very close to each other.

What would be the best description of the dispersion? The standard deviation = 2,186 which can be interpreted that a little less than thirds of the data are within

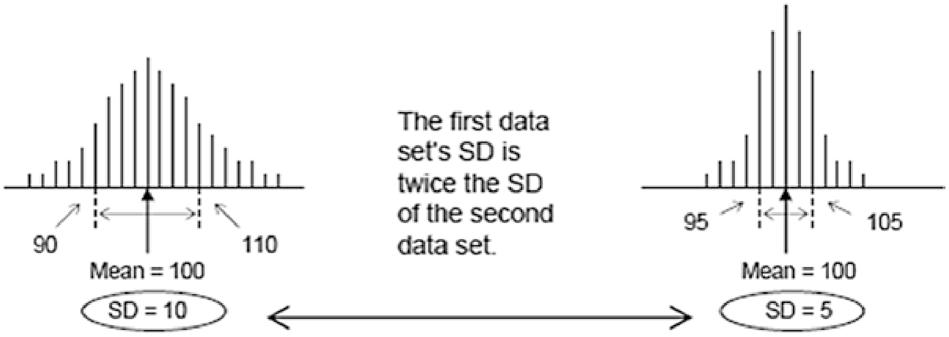

90
110
Mean = 100
SD = 10
The first data set's SD is twice the SD of the second data set.
95
105
Mean = 100
SD = 5

**Fig. 1.15** Two data sets with unequal dispersions and equal means

2,714 ~ 7,086 (i.e., within the interval of mean ± standard deviation) if the contour of the distribution had appeared to a bell-like shape. Because the distribution was not symmetrical, the interquartile range may describe the dispersion better than the standard deviation. The 25th and 75th quartiles are 3,400 and 6,100, respectively, and this tells literally that the half of the group is within this range and the width of the range is 2,700 (i.e., Inter-Quartile Range = 6,100 - 3,400 = 2,700).

### *1.3.8 Dispersions Among Multiple Data Sets*

Figure 1.15 presents two data sets of the same measurement variable in two separate groups of individuals. The two group means are the same but the dispersion of the first group is twice as the dispersion of the second group. The difference in the dispersions is not only visible but is also observed in the standard deviations of 10 and 5.

The comparison of the dispersions may become less straightforward in certain situations. What if the two distributions are from either the same characteristics (e.g., body temperatures) from two distinct groups or different characteristics measured in the same unit but of the same individuals (e.g., fat mass and lean mass in the body measured in grams, or systolic blood pressure (SBP) and diastolic blood pressure measured in mmHg). In Fig. 1.16, can we say the SBP values are more dispersed than DBP solely by reading the two standard deviations? Although the standard deviation of SBP distribution is greater than that of DBP, the mean SBP is obviously also greater and the interpretation of the standard deviations needs to take into account the magnitudes of the two means. Coefficient of Variation (CV) is a descriptive statistic that is applicable for such a circumstance by converting the standard deviation to a universally comparable descriptive statistic.

CV is defined as a standard deviation to mean ratio expressed in percent scale (i.e., CV = 100 × standard deviation/mean). This is useful for comparing the dispersions of two or more distributions of the same variable in two or more different data sets of the means are not identical, or those of two or more different variables measured in the same unit in the same data set. As demonstrated in Table 1.2

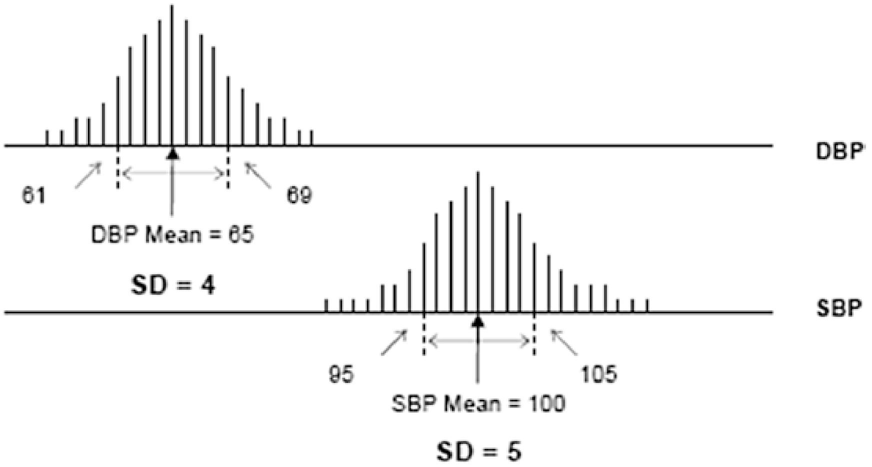


**Fig. 1.16** Two data sets with unequal dispersions and unequal means

**Table 1.2** Application of CV to compare the dispersions of two different characteristics, measured in the same unit, of the same individuals

| | N | Mean | Standard deviation | CV (%) |
|---|---|---|---|---|
| Body fat mass (g) | 160 | 19,783.28 | 8,095.68 | 40.9 |
| Body lean mass (g) | 160 | 57,798.63 | 8,163.56 | 14.1 |

**Table 1.3** Application of CV to compare the dispersions of the same characteristics, measured in the same unit, of two distinct groups

| | | N | Mean | Standard deviation | CV (%) |
|---|---|---|---|---|---|
| Body fat mass (g) | Group 1 | 80 | 21,118.04 | 8,025.78 | 38.0 |
| | Group 2 | 80 | 18,448.53 | 7,993.01 | 43.3 |

demonstrates the situation of comparing the dispersions of two different characteristics measured from the same individuals in the same unit. The standard deviation of the Fat Mass in grams is smaller than that of the Lean Mass in grams of the same 150 individuals, but the CV of the Fat Mass is greater describing that the Fat Mass distribution is more dispersed (CV 43.0 % compared to 14.4 %).

Table 1.3 demonstrates the situation of comparing the dispersions of the same characteristic measured from the same individuals. The standard deviations appeared greater within Group 1 but the CV was greater within Group 2 describing that the dispersion of fat Mass was greater within Group 2.

### *1.3.9 Caution to CV Interpretation*

CV is a useful descriptive statistic to compare dispersions of two or more data sets when the means are different across the data sets. However, the CV should be

**Fig. 1.17** View of Stem-and-Leaf from above

```
Stem-Leaf

 11-2
 10-
 9-3
 8-19
 7-5
 6-115
 5-289
 4-011358
 3-01344577889
 2-68
 1-8
```

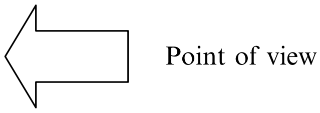
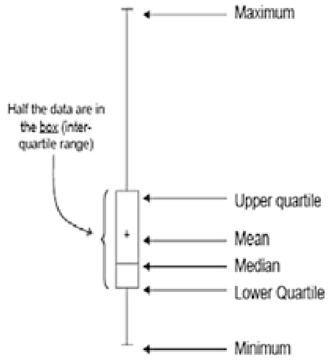


**Fig. 1.18** Relationship between Stem-and-Leaf and Box-and-Whisker plots

```
11-2
10-
9-3
8-19
7-5
6-115
5-289
4-011358
3-01344577889
2-68
1-8
```



applied carefully. When the dispersions of two distributions are compared, we need to ensure that the comparison is appropriate. A comparison of the dispersions of the same or compatible kinds is appropriate (e.g., CVs of body weights obtained from two separate groups, or CVs of SBP and DBP obtained from the same group of persons). However, a comparison of two dispersions of which one of the two is a result of a certain transformation of the original data is not appropriate. For example, in the case of the body temperature example in 1.3.6 the CV of the original °C is 100×(0.62/36.82) = 1.68 % and the CV of the transformed data via °C – 36.5 is 100×(0.62/0.33) = 187.88 %. Did the dispersion increase this large after the whole distribution simple shift? No, the dispersion did not differ and the standard deviations remained the same. However, the CV of °F scale data distribution is different from the original °C scale.

### *1.3.10 Box and Whisker Plot*

Unlike the Stem-and-Leaf plot, this plot does not show the individual data values explicitly. If the Stem-and-Leaf plot is seen from a bird's eye view (Fig. 1.17), then the resulting description can be made as shown in the right hand side panel of Fig. 1.18 which is  depicted separately in Fig. 1.19.

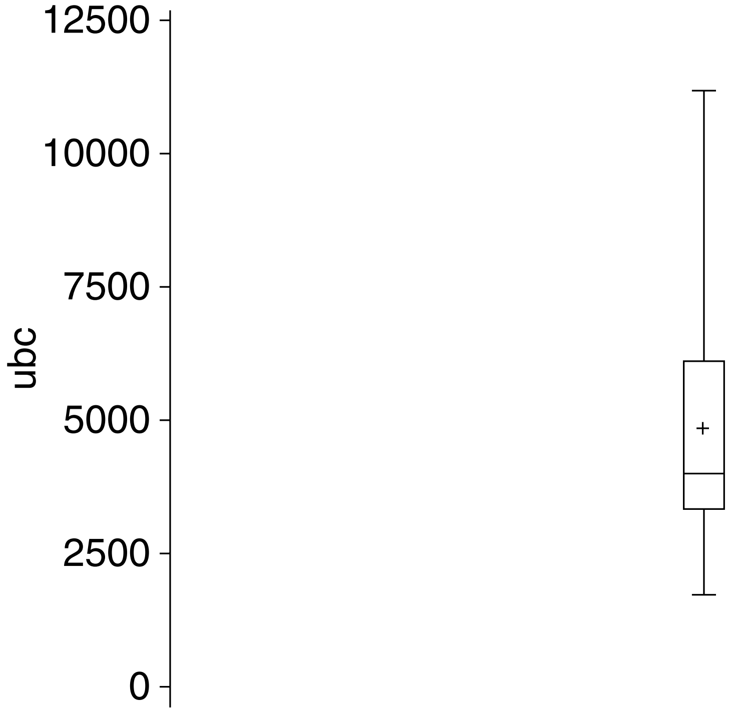


**Fig. 1.19** Box-and-Whisker plot of a skewed data set

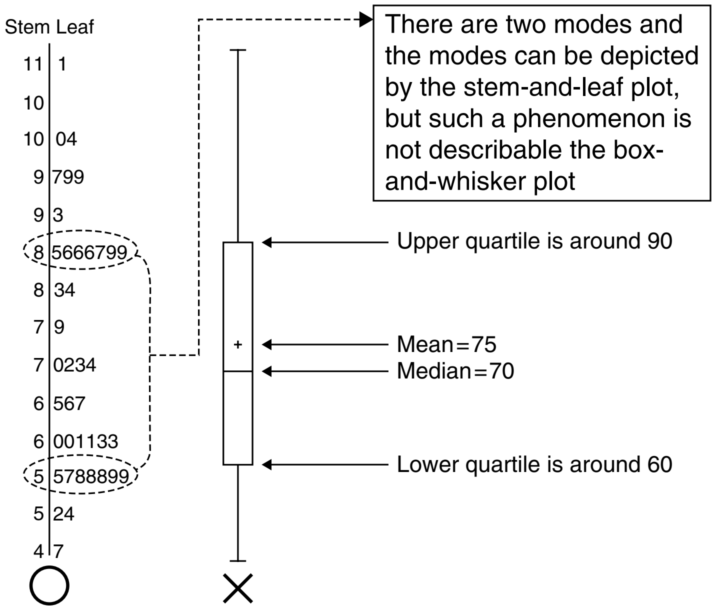


**Fig. 1.20** Stem-and-Leaf and Box-and-Whisker plots of a skewed data set

Among the several advantages of this technique, the unique feature is to visualize the interval where the middle half of the data exist (i.e., the interquartile range) by a box, and the interval where the rest of the data by the whiskers (Fig. 1.19).

If there are two or more modes, the Box-and-Whisker plot cannot fully characterize such a phenomenon, but the Stem-and-Leaf does (see Fig. 1.20).

## 1.4 Descriptive Statistics for Describing Relationships Between Two Outcomes

### *1.4.1 Linear Correlation Between Two Continuous Outcomes*

Previous sections discussed how to summarize the data observed from a single variable (*aka* univariate). This section discusses how to describe a relationship between a set of pairs of continuous outcomes (e.g., a collection of heights measured from biological mother and her daughter pairs). The easiest way to describe such a pattern is to create a scatter plot of the paired data (Fig. 1.21). Correlation coefficient, ρ, is a descriptive statistic that summarizes the direction and strength of a linear association. The correlation coefficient exists between -1 and 1 (geometry of the correlation coefficient is demonstrated by Fig. 1.22). Negative ρ values indicate a reverse linear association between the paired variables and positive ρ values

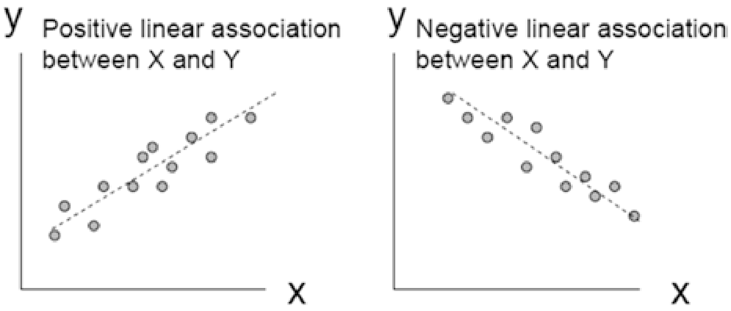
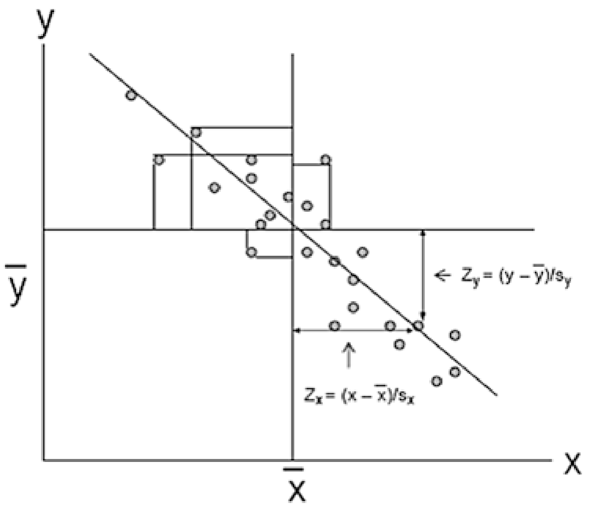


**Fig. 1.21** Linear relationships between two continuous outcomes

Geometry and Calculation of Correlation Coefficient


y
$\bar{y}$
$\leftarrow z_y = (y - \bar{y})/s_y$
↑
$z_x = (x - \bar{x})/s_x$
X
$\bar{x}$

- Measure the standardized horizontal and vertical deviations of individual dots from $(\bar{x}, \bar{y})$, i.e., calculate the $z_x$ and $z_y$ (standardized deviation from the horizontal and vertical means)
- Compute the area of each rectangle and put a sign (Quadrants I and III: + and Quadrants II and IV: -). Note: This is the same as obtaining the product of $z_x$ and $z_y$
- Sum over all signed rectangles, then the resulting value is the correlation coefficient ***ρ***
- The more same "sign" cumulated rectangles, the stronger linear association towards that sign

**Fig. 1.22** Geometry of correlation coefficient

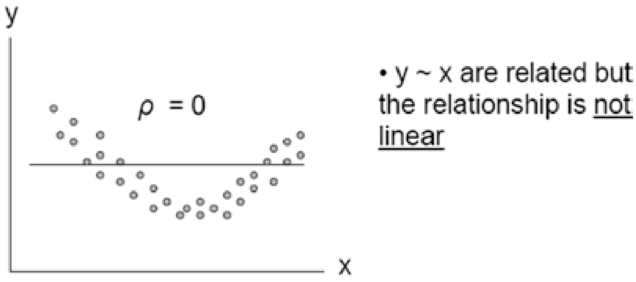


**Fig. 1.23** Nonlinear relationship between two continuous outcomes

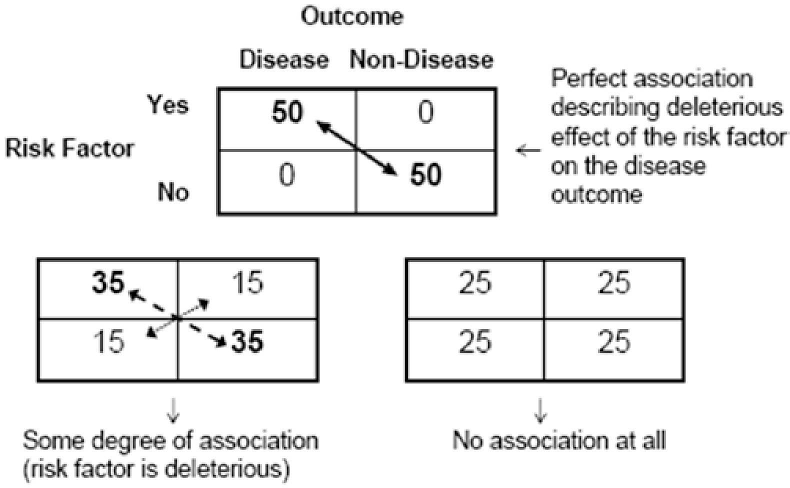
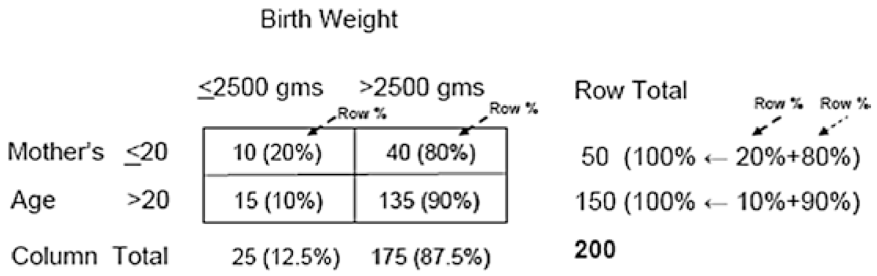

Outcome
Disease
Non-Disease
Risk Factor
Yes
No
50
0
0
50
Perfect association describing deleterious effect of the risk factor on the disease outcome
35
15
15
35
25
25
25
25
Some degree of association (risk factor is deleterious)
No association at all

**Fig. 1.24** Patterns of association between two binary outcomes

indicate the same directional linear association. For example, ρ between *x* and *y*, $\rho_{xy}$ = -0.9 indicates a strong negative linear association between *x* and *y*, and $\rho_{xy}$ = 0.2 indicates a weak positive linear association. Note that the correlation coefficient measures only a linear association. Figure 1.23 illustrates a situation that the correlation coefficient is 0 but there is a clear relationship between the paired variables. The computation may be a burden if done manually. Computer software is widely available, and even Excel can be used (see Chap. 7 for details).

### *1.4.2 Contingency Table to Describe an Association Between Two Categorical Outcomes*

Qualitative categorical outcomes cannot be summarized by the mean and standard deviation value of the observed categories even if the categories were numerically coded (i.e., mean value of such a codified data is meaningless). It is also true that an association of a pair of the numerically categorized outcomes cannot be assessed by the correlation coefficient because the calculation of the correlation coefficient involves the mean value and deviations from the means (see Fig. 1.12). A scatter plot is not well applicable for a visual description between a pair of categorical outcomes. In order to describe the pattern of a set of pairs obtained from two categorical outcomes, the contingency table is used (Fig. 1.24, where each cell number



Birth Weight

| | | ≤2500 gms | >2500 gms | Row Total |
|---|---|---|---|---|
| Mother's | ≤20 | 10 (20%) | 40 (80%) | 50 (100% ← 20%+80%) |
| Age | >20 | 15 (10%) | 135 (90%) | 150 (100% ← 10%+90%) |
| Column Total | | 25 (12.5%) | 175 (87.5%) | 200 |

Row %
Row %
Row %
Row %

**Fig. 1.25** Exploratory data summary by a contingency table

is the observed frequencies of the study subjects). The number appeared in each cell (i.e., cell frequency) provides you the information about the association between two categorical variables. Figure 1.24 illustrates the perfect, moderate, and complete absence of the association between a disease status and a deleterious risk factor. Figure 1.25 illustrates what data pattern is to be recognized for a summary interpretation. There are 20 % (i.e., 10 out of 50) of mothers who are ≤20 years old delivered low weight babies, whereas only 10 % (i.e., 15 out of 150) of the > 20 years old mothers did so. It is also noted that the 20 % is greater than the marginal proportion of the ≤2,500 g (i.e., 12.5 %) and 10 % is lower than the marginal. This observed pattern is interpreted as a twofold difference in proportion of ≥2,500 g between the two mother groups.

### 1.4.3 Odds Ratio

Odds ratio (*OR*) is a descriptive statistic that measures the direction and strength of an association between two binary outcomes. It is defined as a ratio of two odds. The odds is the ratio between the probability of observing an event of interest, π, and the probability of not observing that event, $1-\pi$ (i.e., $odds = \pi/(1-\pi)$). In practical application, the odds can be calculated simply by taking the ratio between the number of events of interest and the number of events not of interest (e.g., number of successes divided by number of failures). Thus the odds ratio associated with a presented risk factor versus the absence of the risk factor for the outcome of interest is defined as $[\pi_1/(1-\pi_1)]/[\pi_2/(1-\pi_2)]$. The odds ratio ranges from 0 to infinity of which the value between 0 and 1 is a protective effect of the factor (i.e., the outcome is less likely to happen within the risk group), 1 being neutral, and greater than 1 is a deleterious effect of the risk factor (i.e., the outcome is less likely to happen within the risk group). According to the definition, the odds ratio associated with the mother's age ≤ 20 years versus > 20 years for the offspring's birth weight ≤ 2,500 g is [0.2/(1-0.2)]/[0.1/(1-0.1)] = 2.25. The same result is obtained simply by the cross product ratio, i.e., [(10/40)]/[(15/135)] = (10 × 135)/(40 × 15) = 2.25. The interpretation of this is that the odds to deliver the offspring with ≤ 2,500 g of birth weight among the mothers age ≤ 20 years is 2.25 times of that of the mothers >20 years. It is a common mistake to make the following erroneous interpretation that the risk of having low birth weight delivery is 2.25 times greater. By definition, the risk is the probability whereas the odds ratio is a ratio of two odds.

## 1.5 Two Useful Probability Distributions

Two important probability distributions are introduced here, which are very instrumental for the inference (see Chap. 2 for inference). A distribution is a complete description of a set of data that species the domain of data occurrences and the corresponding relative frequency over the domain of occurrence. Note that the object being distributed is the relative frequency. A probability model (e.g., Gaussian, binomial model) is the underlying mathematical rule (i.e., mechanism) that generates the data being observed. If you had thought that a distribution is just a curve, or histogram (i.e., visually described data scatter), you would need to revise it.

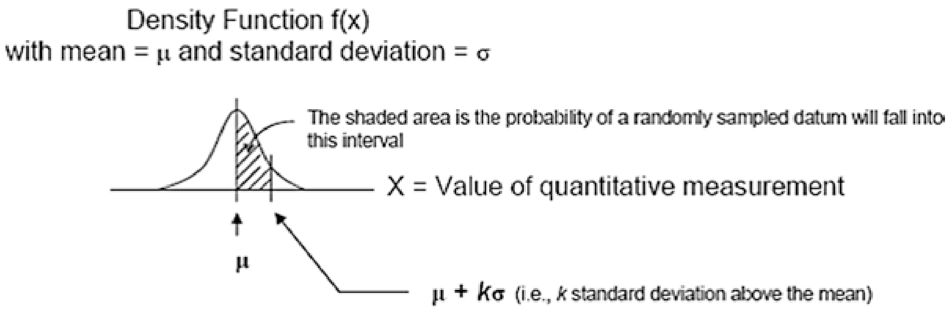
Two widely applied and very useful models in statistical inference are the Gaussian distribution, a continuous data generation mechanism, and binomial distribution, a count of binary event data generation mechanism (i.e., number of presence or absence of a certain characteristic).

### *1.5.1 Gaussian Distribution*

The Gaussian distribution describes the continuous data generation mechanism, and it has important mathematical properties on which the applications of event probability computations and the inference (see Chap. 2) rely. The name Gaussian is originated by the mathematician Gauss who derived its mathematical properties. Its common name is Normal Distribution because the model describes well the probability distributions of typical normal behaviors of continuous outcomes (*aka* bell curve). This distribution has a unique characteristic that the mean, median, and mode are identical, and the data are largely aggregated around the central location and gradually spread symmetrically. A particular Gaussian distribution is completely characterized by the mean and standard deviation, and its notation is $N(\mu, \sigma^2)$, where μ and σ denote the values of mean and standard deviation (thus $\sigma^2$ denotes the variance), respectively.

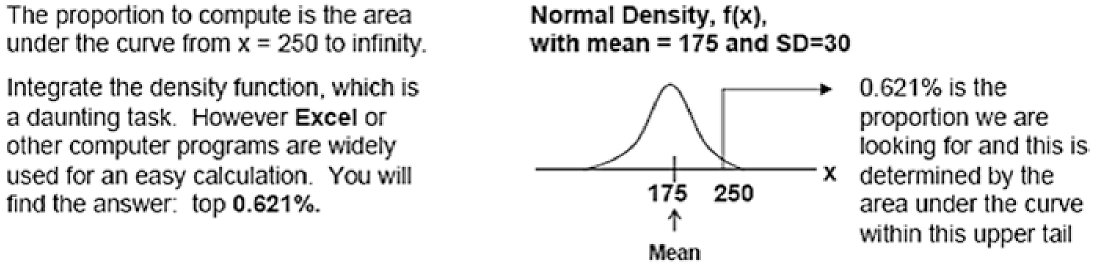
### *1.5.2 Density Function of Gaussian Distribution*

Density is a concentrated quantity on a particular value of the possible data range of a continuous outcome, and this quantity is proportional to the probability of occurrence within a neighborhood of that particular value. Figure 1.26 describes the


Density Function f(x)
with mean = μ and standard deviation = σ
X = Value of quantitative measurement
μ
μ + kσ (i.e., k standard deviation above the mean)

The shaded area is the probability of a randomly sampled datum will fall into this interval

**Fig. 1.26** Gaussian density function curve


The proportion to compute is the area under the curve from x = 250 to infinity.

Integrate the density function, which is a daunting task. However **Excel** or other computer programs are widely used for an easy calculation. You will find the answer: top **0.621%**.

0.621% is the proportion we are looking for and this is determined by the area under the curve within this upper tail

**Fig. 1.27** Density curve and tail probability

density of a Gaussian distribution with mean μ and standard deviation σ. The height of the symmetric bell curve is the size of density (not the actual probability) concentrated over the values of the continuous outcome *x*. The value where the density peaks and the degree of dispersion are completely determined by the mean and standard deviation of the distribution, respectively. The area under the entire density curve becomes 1. As depicted in the figure the shaded area is the probability that the *x* values exist between the mean and *k* times the standard deviation above the mean. The area under the density curve from one standard deviation below to above the mean is approximately 68.3 % (exactly 68.2689 %) meaning that a little bit over middle two-thirds of the group is aggregated symmetrically within one standard deviation around the mean of any Gaussian distribution.

### 1.5.3 Application of Gaussian Distribution

The Gaussian distribution model is very useful tool to approximately calculate a probability of observing certain numerical range of events. The example shown in Fig. 1.27 is to find out the proportion of a large group of pediatric subjects whose serum cholesterol level above 250 mg/mL if the group's cholesterol distribution follows a Gaussian distribution with mean of 175 and standard deviation of 30. Because the standard deviation is 30, the value of 250 is 2.5 times the standard deviation above the mean (i.e., 250 = 175 + 2.5×30). The area under the curve that covers the cholesterol range > 250 is 0.625 %, which indicates the subjects with cholesterol level >250 are within top 0.625 % portion. The calculation requires integration of the Gaussian density function equation. However, we can obtain the result using Excel or standard probability tables of Gaussian distribution. Next section will discuss how to calculate the probability using the tables by transforming any Gaussian distribution to the Standard Normal Distribution.

### *1.5.4 Standard Normal Distribution*

The Standard Normal Distribution is the Gaussian distribution of which the mean is 0 and the standard deviation is 1, i.e., *N (0, 1)*. Any Gaussian distribution can be standardized by the following transformation. In the following equation, $x$ is the variable that represents a value of the original Gaussian distribution with mean μ and standard deviation σ, and $z$ represents the value of the following transformation:

$$z = \frac{x - \mu}{\sigma}$$

This transformation shifts the entire data set uniformly by subtracting μ from all individual values, and rescale the already shifted data values by dividing them by the standard deviation, thus the transformed data will have mean 0 and standard deviation 1.

The Standard Normal Distribution has several useful characteristics on which data analysis and statistical inference rely (we discuss inference well in Chap. 2). First, as seen above, the density is symmetrically distributed over the data range resembling bell-like shape. Moreover, one standard deviation below and above the mean, i.e., the interval from -1 to 1 on $z$, covers approximately 68.3 % of the distribution symmetrically. The interval of $z$ from -2 to 2 (i.e., within two standard deviation symmetrically around the mean) covers approximately 95.5 % of the distribution. The normal range, -1.96 to 1.96 on $z$ which covers 95 % of distribution around mean, is frequently sought (Fig. 1.28).

Figure 1.29, excerpted from Chap. 10, presents the areas under the standard normal density curve covering from negative infinity to various values of the standard normal random variable, $z$. This table can be used to compute the probability evaluated within a certain interval without using a computer program. For example, Pr $\{-1.96 < x \le 1.96\}$ can be computed Pr $\{z \le 1.96\}$ – Pr $\{z \le -1.96\} = 0.975 - 0.025 = 0.95$.

As shown in Fig. 1.30, the probability to observe a value above 250 if the data follow a Gaussian probability model with mean of 175 and standard deviation of 30, then the probability is evaluated by first transforming the value 250 to $z$ value (i.e., standardize to mean 0 and standard deviation 1). The transformed $z$ value is 2.5 (i.e., 250 – 175 = 70, then divide 75 by 30 to find 2.5). Finally, the area under the Standard Normal density curve above 2.5 is the probability of interest. The evaluation of this

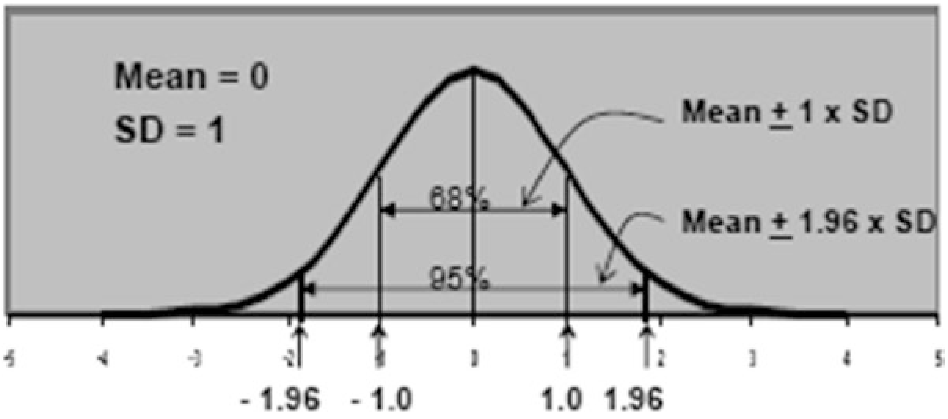
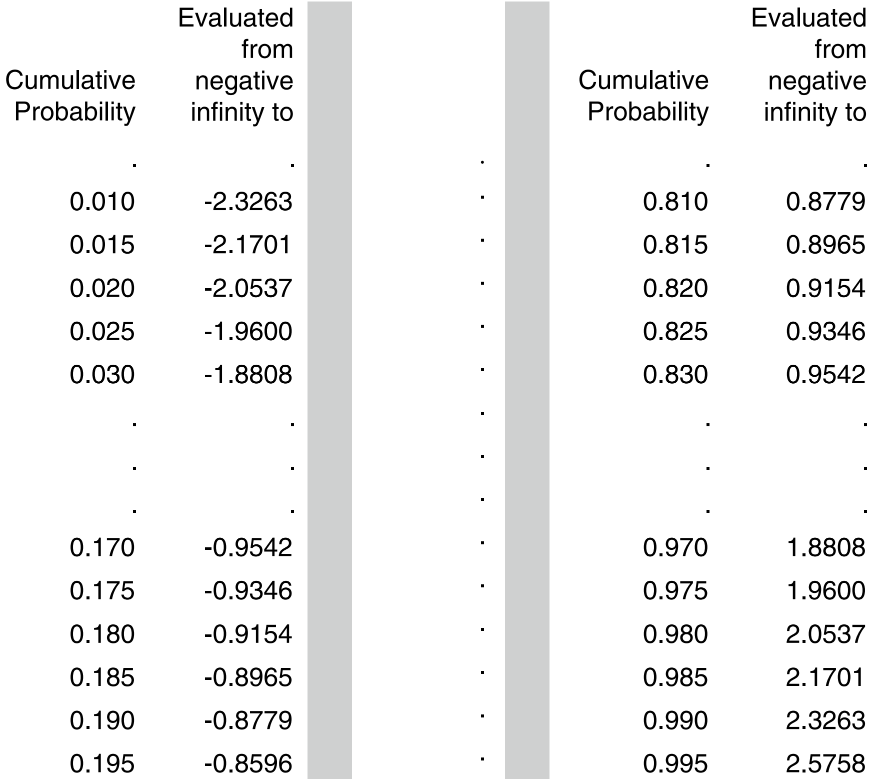


**Fig. 1.28** Covered proportions of 1 (and 1.96) unit of standard deviation above and below means in standard normal distribution



| Cumulative Probability | Evaluated from negative infinity to | | Cumulative Probability | Evaluated from negative infinity to |
|---|---|---|---|---|
| . | . | . | . | . |
| 0.010 | -2.3263 | . | 0.810 | 0.8779 |
| 0.015 | -2.1701 | . | 0.815 | 0.8965 |
| 0.020 | -2.0537 | . | 0.820 | 0.9154 |
| 0.025 | -1.9600 | . | 0.825 | 0.9346 |
| 0.030 | -1.8808 | . | 0.830 | 0.9542 |
| . | . | . | . | . |
| . | . | . | . | . |
| . | . | . | . | . |
| 0.170 | -0.9542 | . | 0.970 | 1.8808 |
| 0.175 | -0.9346 | . | 0.975 | 1.9600 |
| 0.180 | -0.9154 | . | 0.980 | 2.0537 |
| 0.185 | -0.8965 | . | 0.985 | 2.1701 |
| 0.190 | -0.8779 | . | 0.990 | 2.3263 |
| 0.195 | -0.8596 | . | 0.995 | 2.5758 |

**Fig. 1.29** List of selected normal random variates and cumulative probabilities up to those values

area can be done by using either of the tables in Fig. 1.29 or any other published tables. To use the first table, we locate the row of the table associated with $z$ value of -2.5 then narrow down to the first column that lists the calculated area above 2.50 (i.e., 0.9938). If $z$ was 2.53, then the fourth column element of the same row would be read (i.e., 0.9943).

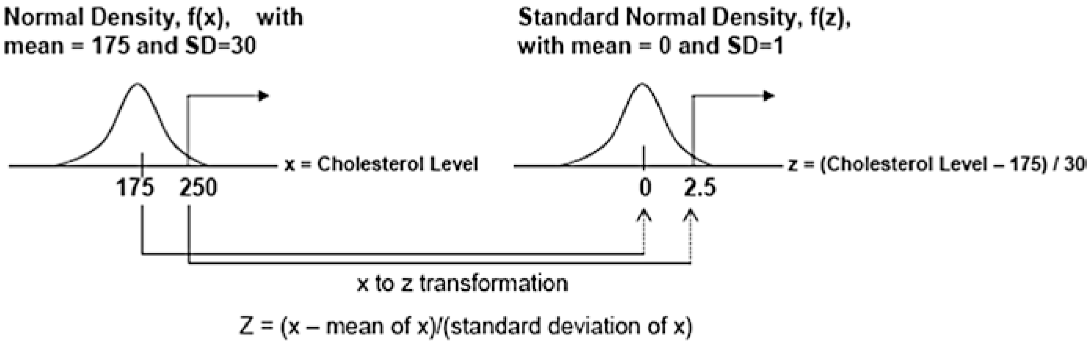

Normal Density, f(x), with
mean = 175 and SD=30
Standard Normal Density, f(z),
with mean = 0 and SD=1
x = Cholesterol Level
175
250
z = (Cholesterol Level – 175) / 30
0
2.5
x to z transformation
Z = (x – mean of x)/(standard deviation of x)

**Fig. 1.30** Standardization of an observed value $x$ = 250 from $N$ (Mean = 175, SD=30) to $z$=2.5 of the standardized normal distribution, i.e., $N$ *(0, 1)*

### *1.5.5 Binomial Distribution*

The probability values that are distributed to the possible numbers of events counted from a set of finite number of dichotomous outcomes (e.g., success and failure) are typically modeled by Binomial Distribution. For a demonstration purpose, let us discuss the following situation. Suppose that it is known that a new investigative therapy can reduce the volume of a certain type of tumor significantly, and the average success rate is 60 %. What will be the probability of observing 4 or more successful outcomes (i.e., significant tumor volume reduction) from a small experiment treating five animals with such a tumor if the 60 % average success rate is true? First, let us calculate the probabilities of all possible outcomes under this assumption, i.e., no success, 1 success, 2, 3, 4, or all 5 successes if the true average success rate is 60 %. Note that a particular subject's single result should not alter the next subject's result, i.e., the resulting outcomes are independent among experimental animals. In this circumstance, the probabilities distributed to the single dichotomous outcome (shrunken tumor as the success or no response as the failure) of each animal are characterized by Bernoulli distribution with its parameter π which is the probability of success in a single animal treatment (i.e., the two probabilities are π, the success rate and 1-$\pi$, the failure rate). The single trial, in this case each trial is a treatment given to each animal, is called Bernoulli trial. The resulting probability distribution of the total number of successes out of those five independent treatment series (i.e., five independent Bernoulli trials) is then described by Binomial Distribution which is characterized by two parameters of which the first is the total number of Bernoulli trials, $n$, and the second is the Bernoulli distribution's parameter of the success rate, π. In this example, the total number of independent trials, $n$, is 5 and the parameter of the success rate, $p$, on each single trial Bernoulli distribution is 0.6. Table 1.4 lists all possible results and their probabilities (0 = failure with its single occurring chance of 0.4, 1=success with its single occurring chance of 0.6). As shown in the last column of the table, these computed probabilities are 0.0102 for 0 successes (i.e., all failures and its probability is 0.4 × 0.4 × 0.4 × 0.4 × 0.4 = 0.0102), 0.0768 for 1 success, 0.2304 for 2 successes, 0.3456 for 3 successes,

**Table 1.4** *Bi (5, 0.6),* binomial distribution with n=5 and **π**=0.6

| Number of successes | Result of subjects | | | | | Probability |
|---|---|---|---|---|---|---|
| | 1st | 2nd | 3rd | 4th | 5th | |
| 0 (1 assortment) | 0 | 0 | 0 | 0 | 0 | 0.4 × 0.4 × 0.4 × 0.4 × 0.4 = $0.4^{5}$ |
| | | | | | | (Subtotal = 0.0102) |
| 1 (5 assortments) | 1 | 0 | 0 | 0 | 0 | **0.6** × 0.4 × 0.4 × 0.4 × 0.4 = 0.6 × $0.4^{4}$ |
| | 0 | 1 | 0 | 0 | 0 | 0.4 × **0.6** × 0.4 × 0.4 × 0.4 = 0.6 × $0.4^{4}$ |
| | 0 | 0 | 1 | 0 | 0 | 0.4 × 0.4 × **0.6** × 0.4 × 0.4 = 0.6 × $0.4^{4}$ |
| | 0 | 0 | 0 | 1 | 0 | 0.4 × 0.4 × 0.4 × **0.6** × 0.4 = 0.6 × $0.4^{4}$ |
| | 0 | 0 | 0 | 0 | 1 | 0.4 × 0.4 × 0.4 × 0.4 × **0.6** = 0.6 × $0.4^{4}$ |
| | | | | | | (Subtotal = 0.0768) |
| 2 (10 assortments) | 1 | 1 | 0 | 0 | 0 | **0.6** × **0.6** × 0.4 × 0.4 × 0.4 = $0.6^{2}$ × $0.4^{3}$ |
| | 1 | 0 | 1 | 0 | 0 | **0.6** × 0.4 × **0.6** × 0.4 × 0.4 = $0.6^{2}$ × $0.4^{3}$ |
| | | | | | | (Subtotal = 0.2304) |
| 3 (10 assortments) | 1 | 1 | 1 | 0 | 0 | **0.6** × **0.6** × **0.6** × 0.4 × 0.4 = $0.6^{3}$ × $0.4^{2}$ |
| | 1 | 1 | 0 | 1 | 0 | **0.6** × **0.6** × 0.4 × **0.6** × 0.4 = $0.6^{3}$ × $0.4^{2}$ |
| | | | | | | (Subtotal = 0.3456) |
| 4 (5 assortments) | 1 | 1 | 1 | 1 | 0 | **0.6** × **0.6** × **0.6** × **0.6** × 0.4 = $0.6^{4}$ × 0.4 |
| | 1 | 0 | 1 | 1 | 1 | **0.6** × 0.4 × **0.6** × **0.6** × **0.6** = $0.6^{4}$ × 0.4 |
| | | | | | | (Subtotal = 0.2592) |
| 5 (1 assortment) | 1 | 1 | 1 | 1 | 1 | **0.6** × **0.6** × **0.6** × **0.6** × **0.6** = $0.6^{5}$ |
| | | | | | | (Subtotal = 0.0778) |

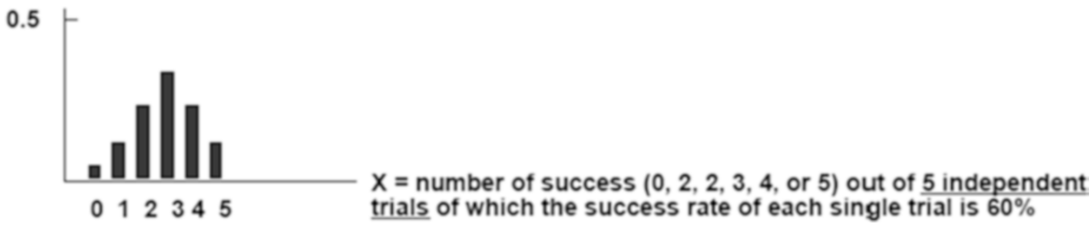
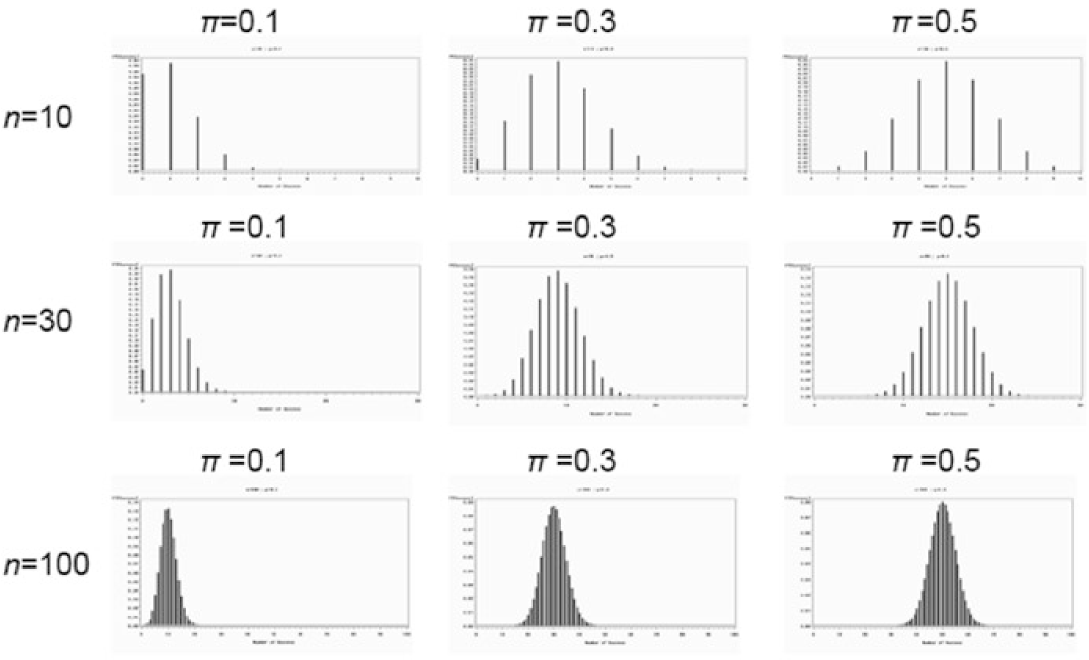


**Fig. 1.31** Distribution (*aka* probability mass function) of *Bi* (n=5, π=0.6)

0.2592 for 4 successes, and 0.0778 for all 5 successes, respectively. General notation of a Binomial Distribution is *Bi (n, π)*, thus in this example it is *Bi (5, 0.6)*. Let us also note that the aforementioned Bernoulli distribution is a special case of Binomial Distribution, and its general notation is *Bi (1, π)*. Figure 1.31 displays *Bi (5, 0.6)*. Thus the probability of observing 4 or more successes out of the treatments given to five independent animals is 0.2592 + 0.0778 = 0.3370. Although this book does not exhibit the closed form equation that completely describes the Binomial Distribution, the following expression can help understand the concept: *Bi (n, π)* can be expressed by *Probability of {no. of events, X = x out of n independent Bernoulli trials} = K* $\pi^{x}(1\text{-}\pi)^{n-x}$, where *K* is an integer value multiplier that reflects the number all possible assortments of the number of success events *x* (*x* = 0, 1, …, *n*). Readers who are familiar with combinatorics can easily figure out *K = n!/[x!(n-x)!]*. In Table 1.4, *K* =1 for *x* = 0, *K* = *5* for *x* = 1, *K* = 10 for *x*= 2, …, and *K* = 1 for



Note: As *n* for a given *π* becomes large, or *π* becomes large for a given *n* the Binomial distribution becomes closer to a normal distribution with mean = *nπ* and variance = *nπ* (1- *π* ).

**Fig. 1.32** Large sample behavior of binomial distributions illustrated by histograms of binomial distributions with various trial sizes and success rates

$x = 5$. It is straightforward that the expression of *Bi (1, π)* is *Probability of {no. of events, X = x out of 1 Bernoulli trials}* $= \pi^x(1\text{-}\pi)^{1\text{-}x}$, where $x$ = either 1 (for success) or 0 (failure).

While the Binomial Distribution fits to the probability of success counts arising from a fixed number of independent trials, if the event of interest is not rare (i.e., π is not very small) and the size of the trial, *n*, becomes large, then the probability calculation for a range of number of success events can be conveniently approximated by using the Gaussian distribution even if, the number of success is not continuous. Figure 1.32 demonstrates the rationale for such an application. In general, for $n \times \pi \geq 5$ (i.e., *the number of expected successes is at least 5),* if *n* becomes large for a given a π, or π becomes large for a given *n,* then the distributed probability pattern of Binomial Distribution becomes closer to $N\,(\mu = n \times \pi,\ \sigma^2 = n \times \pi \times (1\text{-}\ \pi))$.

Suppose that we now increased the number of animal experiment to 100, and we want to compute the probability of observing 50–75 successes arising from 100 independent trials. Because $n \times \pi = 100 \times 0.6 = 60$, and $n \times \pi \times (1\text{-}\ \pi) = 100 \times 0.6 \times 0.4 = 24$, this task can be resorted to the normal approximation for which the used distribution is $N\,(\mu = 60,\ \sigma^2 = 24)$. Then as depicted by Fig. 1.33, the first step is to transform the interval 50 ~ 75 on $\underline{N\,(\mu = 60,\ \sigma^2 = 24)}$ to a new interval on $N\,(0,\ \underline{1}),$ i.e., 50 → $(50 - \mu)/\sigma = (50\text{-}60)/\sqrt{24} = \text{-}2.05$ and 75 → $(70 - \mu)/\sigma = (75\text{-}60)/\sqrt{24} = 2.05$. So, the probability to observe 50–75 successes is the area under the density curve of $N\,(0,\ 1)$ covering from -2.05 and 2.05 on *z,* which is 0.98.

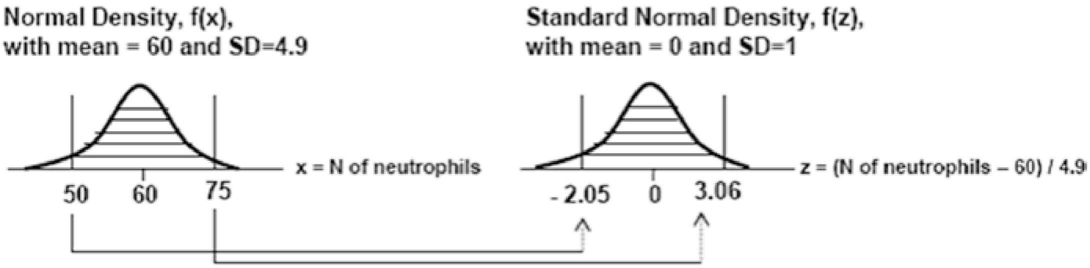

Normal Density, f(x),
with mean = 60 and SD=4.9
Standard Normal Density, f(z),
with mean = 0 and SD=1
x = N of neutrophils
z = (N of neutrophils – 60) / 4.9
50
60
75
- 2.05
0
3.06

**Fig. 1.33** Normal approximation to calculate a probability range of number of binary events

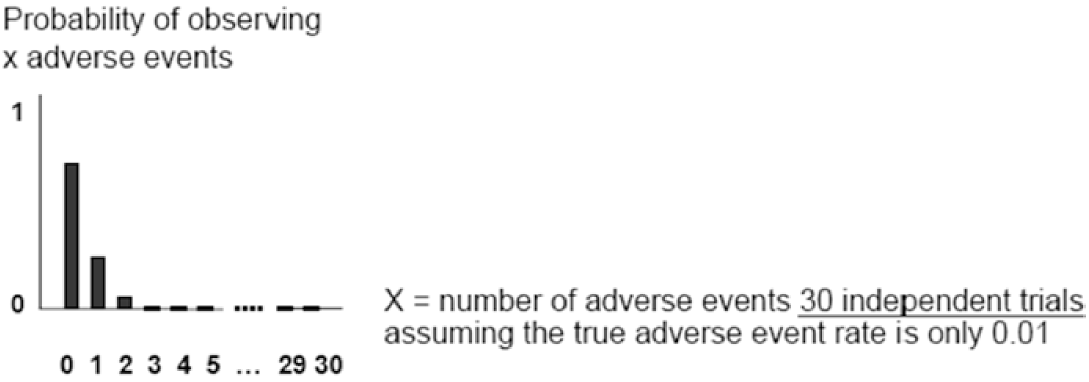


**Fig. 1.34** Distribution (*aka* probability mass function) of *Bi* (n=30, **π** =0.01)

On the other hand, when the event of interest is rare and the size of the trial becomes very large then the computation can be approximated by Poisson model in which the number of trials is no longer an important constant (i.e., parameter) that characterizes the Poisson distribution. The notation is *Poi (λ),* where λ denotes the number of average successes of the rare event out of a large number of independent trials. A particular exemplary outcome that is well characterized by the Poisson model is the number of auto accidents on a particular day in a large metropolitan city. The rare events can be the ones of which the Binomial characteristic constants are $n \times \pi < 5$ (i.e.*, expected number of successes).* The next example is a Binomial Distribution for which the probability calculation can be approximated by a Poisson distribution. Figure 1.34 displays the probabilities of observing 0, 1, 2, …, 30 adverse events among 30 independent clinical trials of a new drug if the true adverse event rate = 0.01 (i.e., 1 %). The typical pattern of Poisson distribution is that the probability value decreases exponentially after certain number of successes, and as the expected number of successes, $n \times \pi$, becomes smaller the value decreases faster. If we let a computer calculate the probability to observe 3 or more adverse events from 30 trials, then the result will be 0.0033. If we approximate this distribution to Poi ($\lambda = 30 \times 0.01 = 0.3$) and let a computer calculate such an event, the result will be 0.0035, which is not much different from the Binomial model-based calculation.

## 1.6 Study Questions

1. What are the similarity and dissimilarity between the interval scale and ratio scale?
2. What is the definition of a distribution? What is being distributed?
3. In a Box-and-Whisker plot, what proportion of the population is contained in the "box" interval? Is such a plot useful to describe a bimodal (i.e., two modes) distribution?
4. Please explain the definition of standard deviation.
5. What proportion of the data values are within one standard deviation above and below the mean if the data are normally distributed?
6. Can a correlation coefficient measure the strength of any relationship between two continuous observations?
7. What are the definitions of odds and odds ratio?
8. What are the two parameters that completely determine a Gaussian distribution?
9. What are the two parameters that completely determine a Binomial Distribution?
10. Under what condition can a Gaussian model approximate the proportion of a population lies within a certain range of number of events describable by a Binomial model?

## Bibliography

Grimmett G, Stirzaker D (2001) Probability and random processes, 3rd edn. Oxford University Press, Oxford

Johnson NL, Kotz S, Balakrishnan N (1994) Continuous univariate distributions, vol 1, 2nd edn. John Wiley, New York

Ross S (2010) A first course in probability, 8th edn. Pearson Prentice Hall, Upper Saddle River

Snecdecor GW, Cochran WG (1991) Statistical methods, 8th edn. Wiley-Blackwell, Oxford

Tukey JW (1977) Exploratory data analysis. Addison-Wesley, New York