1

Statistical measures

1.1 Introduction

Medical professionals, hospitals and healthcare centers record heights, weights and other relevant physical measurements of patients along with their blood pressures cholesterol levels and similar diagnostic measurements. National organizations such as the Center for Disease Control (CDC) in the United States, the World Health Organization (WHO) and several national and international organizations record and analyze various aspects of the healthcare status of the citizens of all age groups. Epidemiological studies and surveys collect and analyze health-related information of the people around the globe. Clinical trials and experiments are conducted for the development of effective and improved medical treatments.

Statistical measures are utilized to analyze the various diagnostic measurements as well as the outcomes of clinical experiments. The mean, mode and median described in the following sections locate the centers of the distributions of the above types of observations. The variance, standard deviation (S.D.) and the related coefficient of variation (C.V.) are the measures of dispersion of a set of observations. The quartiles, deciles and percentiles divide the data respectively into four, ten and one hundred equal parts. The skewness coefficient exhibits the departure of the data from its symmetry, and the kurtosis coefficient its peakedness. The measurements on the heights, weights and Body Mass Indexes (BMIs) of a sample of twenty-year-old boys obtained from the Chart Tables of the CDC (2008) are presented in Table 1.1. These measurements for the ten and sixteen- year old boys and girls are presented in Appendix Tables T1.1–T1.4.

Statistical Methodologies with Medical Applications, First Edition. Poduri S.R.S. Rao.
© 2017 John Wiley & Sons, Ltd. Published 2017 by John Wiley & Sons, Ltd.

Table 1.1 Heights (cm), weights (kg) and BMIs of twenty-year old boys.

Height	Weight	BMI
162	54	20.58
163	55	20.70
167	58	20.80
168	59	20.90
170	60	20.76
172	62	20.96
172	63	21.30
173	66	22.05
174	68	22.46
176	72	23.24
176	75	24.21
176	75	24.21
177	78	24.90
178	80	25.25
178	82	25.88
180	84	25.93
184	86	25.40
184	88	25.99
186	95	27.46
188	102	28.86

BMI = Weight/(Height)^2.

1.2 Mean, mode and median

The diagnostic measurements of a sample of n individuals can be represented by $x_i, i = (1,2,\ldots,n)$. Their mean or average is

$$\bar{x} = \sum_{i=1}^{n} x_i/n = (\mathrm{x}_1 + \mathrm{x}_2 + \ldots + \mathrm{x}_n)/n. \tag{1.1}$$

For the heights of the boys in Table 1.1, the mean becomes $\bar{x} = (162 + 163 + \ldots + 188)/20 = 175.2$ cm. Similarly, the mean of their weights is 73.1 kg. For the BMI, which is (Weight/Height^2), the mean becomes 23.59.

The mode is the observation occurring more frequently than the remaining observations. For the heights of the boys, it is 176 cm. The median is the middle value of the observations. If the number of observations n is odd, it is the $(n + 1)$th observation. If n is an even number, it is the average of the $(n/2)$th and the next observation. Both the mode and median of the twenty heights of the boys in Table 1.1 are equal to 176 cm, which is slightly larger than the mean of 175.2 cm.

2	16	23
4	16	78
9	17	02234
(6)	17	666788
5	18	044
2	18	68

Figure 1.1 Stem and leaf display of the heights of the twenty boys. Leaf unit = 1.0. The median class has (6) observations. The cumulative number of observations below and above the median class are (2, 4, 9) and (5, 2).

The mean, mode and median locate the center of the observations. The mean is also known as the first moment m_1 of the observations. For the healthcare policies, for instance, it is of importance to examine the average amount of the medical expenditures incurred by families of different sizes or specified ranges of income. At the same time, useful information is provided by the median and modal values of their expenditures. Figure 1.1 is the Stem and Leaf display of the heights in Table 1.1. The cumulative number of observations below and above the median appear in the first column. The second and third columns are the stems, with the attached leaves.

1.3 Variance and standard deviation

The variance is a measure of the dispersion among the observations, and it is given by

$$\begin{aligned} s^2 &= \sum_{i=1}^{n}(x_i-\bar{x})^2/(n-1) \\ &= \left[(x_i-\bar{x})^2+(x_2-\bar{x})^2+\ldots(x_n-\bar{x})^2\right]/(n-1). \end{aligned} \tag{1.2}$$

The divisor $(n-1)$ in this expression represents the *degrees of freedom* (d.f.). If $(n-1)$ of the observations and the sum or mean of the n observations are known, the remaining observation is automatically determined. The expression in (1.2) can also be expressed as $\sum_{i\neq j}\left(x_i-x_j\right)^2/n(n-1)$, which is the average of the squared differences of the $n(n-1)$ pairs of the observations. The standard deviation (S.D.) is given by **s**, the positive square root of the variance. The second central moment of the observations $m_2=\sum(x_i-\bar{x})^2/n$ is the same as $(n-1)s^2/n$. For the twenty heights of boys in

Table 1.2 Summary figures for the heights, weights and BMIs of the 20 boys in Table 1.1.

	Height	Weight	BMI
Mean ($\bar{x}$)	175.2	73.1	23.59
Variance (s^2)	51.33	188.09	6.53
m_2	48.76	178.69	6.21
S.D.(s)	7.16	13.71	2.56
C.V.(%)	4.09	18.76	10.85
m_3	–18.86	913.69	5.24
K_1	–0.055	0.383	0.341
m_4	5690	70901	75.93
K_2	2.39	2.22	1.97

Table 1.1, $s^2 = \left[(162-175.2)^2 + (163-175.2)^2 + \ldots + (188-175.2)^2\right]/19 = 51.33$ and $m_2 = (19/20)(51.33) = 48.76$. The standard deviation becomes $s = 7.16$ cm.

The unit of measurement is attached to both the mean and standard deviation; kg for weight and cm for height. It is kg/(meter-squared) for the BMI. The coefficient of variation (C.V.), is the ratio of the standard deviation to the mean $(s/\bar{x})$ and is devoid of the unit of measurement of the observations. The mean, variance, standard deviation and C.V. for the above three characteristics for the 20 boys in Table 1.1 are presented Table 1.2.

1.4 Quartiles, deciles and percentiles

Any set of data can be arranged in an ascending order and divided into four parts with one quarter of the observations in each part. Twenty-five percent of the observations are below the first quartile Q_1 and 75 percent above. Similarly, half the number of observations are below the median, which is the second quartile Q_2, and half above. Three-quarters of the observations are below the third quartile Q_3 and one-fourth above. As seen in Section 1.2, the median of the heights in Table 1.1 is 176 cm. The average of the fifth and sixth observations is 171 cm, which is the first quartile. Similarly, the third quartile is 179 cm, which is the average of the fifteenth and sixteenth observations. The box and whiskers plot in Figure 1.2 presents the positions of these quartiles.

Ten percent of the observations are below the first decile and 90 percent above. Ninety percent of the observations are below the ninth decile and 10 percent above. One percent of the observations are below the first percentile and 99 percent above. Similarly, 99 percent of the observations are below the ninety-ninth percentile and 1 percent above.

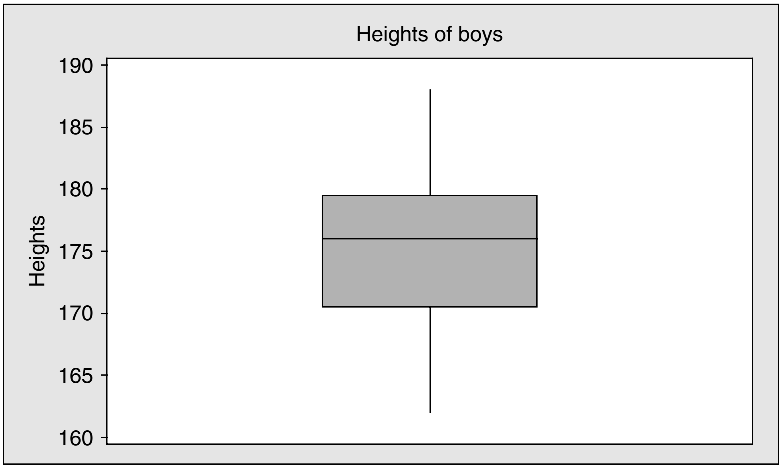

Figure 1.2 Box and whiskers plot of the heights of boys in Table 1.1, obtained from Minitab. *The middle line of the box is the median Q_2. The bottom and top lines are the first and third quartiles Q_1 and Q_3. The tips of the vertical line, whiskers, are the upper and lower limits $Q_1 + 1.5(Q_3 - Q_1)$ and $Q_1 - 1.5(Q_3 - Q_1)$.*

1.5 Skewness and kurtosis

Physical or diagnostic measurements $x_i, i = (1, 2, \ldots, n)$, of a group of individuals may not be symmetrically distributed about their mean. The third central moment, $m_3 = \sum_1^n (x_i - \bar{x})^3/n$ will be zero if the observations are symmetrically distributed about the mean. It will be positive if the observations are skewed to the right and negative if they are skewed to the left. For the symmetrically distributed observations, the third, fifth, seventh and all the odd central moments will be zero. The *Pearsonian coefficient of skewness* is given by $\mathrm{K}_1 = m_3/m_2^{3/2}$, which does not depend on the unit of measurement of the observations unlike m_2 and m_3. For any set of observations symmetrically distributed about its mean, $m_3 = 0$ and hence $\mathrm{K}_1 = 0$. For the positively skewed observations, m_3 and K_1 are positive. For the negatively skewed observations, they are negative. For the heights of the boys in Table 1.1, $m_3 = -18.86$ and $\mathrm{K}_1 = -18.86/\,(48.76)^{3/2} = -0.055$. These heights are slightly negatively skewed.

The fourth central moment of the observations, $m_4 = \sum_1 (x_i - \bar{x})^4/n$, becomes large as the distribution of the observations becomes peaked and small as it becomes flat. The *Pearsonian coefficient of kurtosis* is given by $\mathrm{K}_2 = m_4/m_2^2$, which does not depend on the unit of measurement. For the normal distribution, which is extensively employed for statistical analysis and inference, $\mathrm{K}_1 = 0$ and $\mathrm{K}_2 = 3$. For the observations on all the three characteristics in Table 1.1, the fourth moments are large, as seen from Table 1.2, but K_2 is smaller than three.

1.6 Frequency distributions

Any set of clinical measurements or medical observations can be classified into a convenient number of groups and presented as the *frequency distribution*. The CDC, National Center for Health Statistics (NCHS) and other organizations present various health-related measurements on the U.S. population in the form of summary tables. These measurements are obtained from periodic or continual surveys of the population in the country and also from the administrative medical records of the population. They are arranged according to age groups, education, income levels, male-female classification and other characteristics of interest. Similar summary figures are presented by the WHO and healthcare organizations throughout the world. For the sake of illustration, the twenty heights of the boys in Table 1.1 are arranged in Table 1.3 into seven classes of the same width of five, and displayed as the histogram in Figure 1.3.

In general, the n observations can be divided into k classes with n_i observations in the ith class, $n = \sum_1^k n_i$. The mid-values of the classes can be denoted by $(x_1, x_2, \ldots, x_\mathrm{k})$. With the above notation, the mean of the n observations becomes

$$\bar{x} = \sum_1^k f_i x_i = (n_1 x_1 + n_2 x_2 + \ldots + n_k x_k)/n, \tag{1.3}$$

where $f_i = n_i/n$ is the relative frequency in the ith class and $\sum_1^k f_i = 1$. From the above table and (1.3), the mean of the heights is

$$\bar{x} = (1 \times 160 + 2 \times 165 + \ldots + 1 \times 190)/20 = 175.25.$$

Since the 20 observations are grouped, this mean differs slightly from the actual value of 175.2 cm.

Table 1.3 Frequency distribution of the heights of the 20 boys in Table 1.1.

Class	Mid-x_i	Frequency (n_i)	Relative frequency (f_i)
157.5–162.5	160	1	0.05
162.5–167.5	165	2	0.10
167.5–172.5	170	4	0.20
172.5–177.5	175	6	0.30
177.5–182.5	180	3	0.15
182.5–187.5	185	3	0.15
187.5–192.5	190	1	0.05
		$n = 20$	$\Sigma f_i = 1$

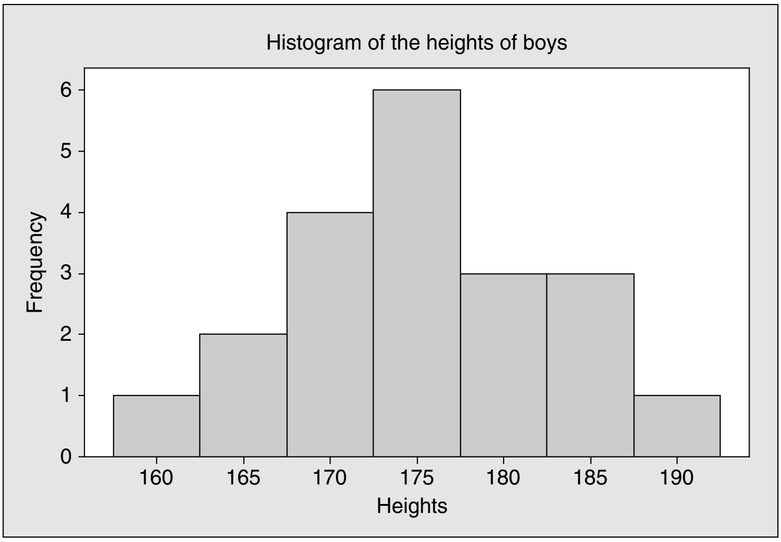

Figure 1.3 Histogram of the distribution of the heights of the boys in Table 1.3 obtained from Minitab.

For the grouped data, the second moment becomes

$$m_2 = \sum_{i=1}^{k} f_i (x_i - \bar{x})^2 = \left[n_1 (x_1 - \bar{x})^2 + n_2 (x_2 - \bar{x})^2 + \ldots + n_k (x_k - \bar{x})^2 \right] / n. \tag{1.4}$$

Now, $s^2 = \sum_{1}^{k} n_i (x_i - \bar{x})^2 / (n-1) = n m_2 / (n-1)$. From (1.4), for the heights of the boys, $m_2 = 56.19$ and $s^2 = 59.14$, which differ from the actual values 48.76 and 51.33 as a result of the grouping. From the grouped data, the third and fourth central moments are obtained from $m_3 = \sum_{1}^{k} f_i (x_i - \bar{x})^3$ and $m_4 = \sum_{1}^{k} f_i (x_i - \bar{x})^4$. In general, the rth central moment for the grouped data is given by $m_r = \sum_{1}^{k} f_i (x_i - \bar{x})^r$.

1.7 Covariance and correlation

The heights and weights of the 20 boys in Table 1.1 can be denoted by (x_i, y_i), $i = (1, 2, \ldots, n)$. With the subscripts *(x, y)* for these characteristics, as presented in Table 1.2, the standard deviations of these characteristics are $s_x = 7.16$ and $s_y = 13.71$. Their covariance is given by

$$s_{xy} = \sum_{1}^{n} (x_i - \bar{x})(y_i - \bar{y})/(n-1)$$

$$= [(x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \ldots + (x_n - \bar{x})(y_n - \bar{y})]/(n-1). \qquad (1.5)$$

It is the sum of the cross-products of the deviations of (x_i, y_i) from their means divided by (n – 1). It can also be expressed as $s_{xy} = \left(\sum_{1}^{n} x_i y_i - n\bar{x}\bar{y} \right) / (n-1)$. The sample correlation coefficient of (x, y) is

$$r = s_{xy}/s_x s_y. \qquad (1.6)$$

It will be positive as y increases with x and negative if it decreases, and vice versa. In general, the covariance can be positive or negative. It can range from a very small negative value to a very large positive number, and the units of measurements of both x and y are attached to it. The correlation coefficient, however, ranges from –1 to + 1, and it is devoid of the units of measurements of the two characteristics. If x increases as y increases, or x decreases as y decreases, their covariance and correlation will be positive; negative otherwise. If x and y are not related, s_{xy} and r will be zero. For the heights and weights of the twenty-year-old boys in Table 1.1, from (1.5), (1.6) and Table 1.2, $s_{xy} = (1814.61/19) = 95.51$ and $r = 95.51/(7.16 \times 13.71) = 0.97$. In this case, these two characteristics are highly positively correlated as expected. Figure 1.4 displays the relationship of the weights and heights of the twenty boys in Table 1.1.

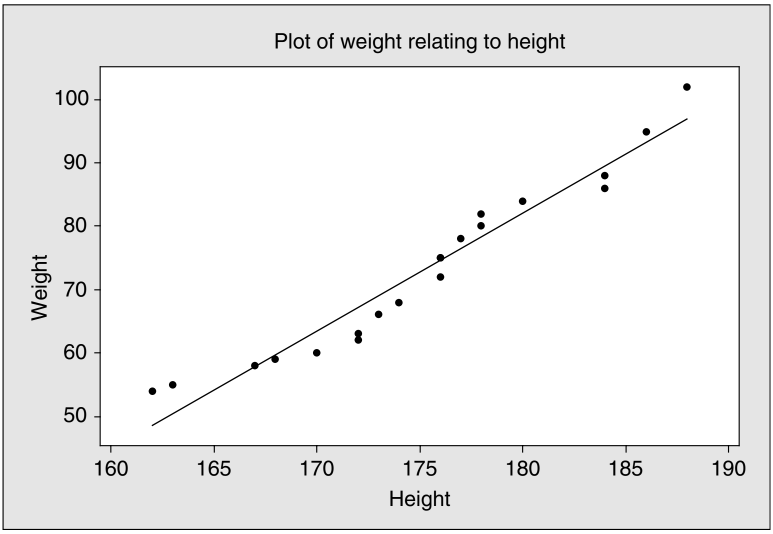

Figure 1.4 Plot of the weights of the twenty-year-old boys on their heights from the observations in Table 1.1.

1.8 Joint frequency distribution

When the number of observations on two variables (x, y) is not small, they can be grouped into the joint frequency distribution. National and international organizations present the health-related characteristics in this form. For the sake of illustration, age (x) and weight (y) of a sample of $n = 200$ adults classified into $r = 3$ rows and $c = 3$ columns are presented in Table 1.4.

With the first and second subscripts $i = (1, 2, \ldots, r)$ and $j = (1, 2, \ldots, c)$ representing the rows and columns respectively, the ith row and jth column consists of n_{ij} adults. The total number of observations in the ith row and jth column respectively are $n_{i.} = \sum_{j=1}^{c} n_{ij}$ and $n_{.j} = \sum_{j=1}^{r} n_{ij}$. The overall sample size becomes $n = \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij} = \sum_{i=1}^{r} n_{i.} = \sum_{j=1}^{c} n_{.j}$. The row and column totals are the marginal totals. They provide the frequency distributions of the age and weight respectively. The means, variances and standard deviations for the row and column classifications are obtained from these distributions as described in Section 1.6. With the mid-values $(x_1, x_2, \ldots, x_{\mathrm{r}})$ of the row classification and $(y_1, y_2, \ldots, y_{\mathrm{c}})$ of the column classification, the covariance of (x, y) is obtained from

$$s_{xy} = \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij}(x_i - \bar{x})(y_i - \bar{y})/(n-1). \tag{1.7}$$

The correlation coefficient is found from $r = s_{xy}/s_x s_y$.

From Table 1.4, the mean, variance and standard deviation of the age are

$$\bar{x} = (35 \times 70 + 45 \times 90 + 55 \times 40)/200 = 43.5,$$

$$s_x^2 = \left[70(35-43.5)^2 + 90(45-43.5)^2 + 40(55-43.5)^2\right]/199 = 53.02,$$

and $s_x = 7.28$.

Table 1.4 Age (x) and weight (y) of $n = 200$ adults.

		Weight (lbs)			
		140–150	150–160	160–170	Total
Age (years)	30–40	20	30	20	70
	40–50	15	35	40	90
	50–60	5	15	20	40
	Total	40	80	80	200

Similarly, for the weight, $\bar{y} = 157$, $s_y^2 = 56.28$ and $s_y = 7.50$. From (1.7),

$$s_{xy} = [20(35-43.5)(145-157) + \ldots + 20(55-43.5)(165-157)]/199 = 10.55.$$

The correlation of age and weight now becomes $r_{xy} = 10.55/(7.28 \times 7.50) = 0.19$, which is not very high.

1.9 Linear transformation of the observations

For computations, it may become convenient to transform the data first. For instance, we may subtract 170 from each of the heights in Table 1.1, and divide the result by 10. The new observations now become $u_i = (x_i - 170)/10 = (1/10)\,x_i - 17$. We may also first divide each height by 100 and then subtract 5. Now, $u_i = (1/100)x_i - 5$. In either case, the new observations take the form of $u_i = ax_i + b$, where (a, b) are positive or negative constants. The mean of the transformed observations becomes

$$\bar{u} = \sum u_i/n = \sum (ax_i + b)/n = a\sum x_i/n + b = a\bar{x} + b. \tag{1.8}$$

Their variance becomes

$$\begin{aligned} s_u^2 &= \sum (u_i - \bar{u})^2/(n-1) \\ &= \sum [(ax_i + b) - (a\bar{x} + b)]^2/(n-1) \\ &= a^2 \sum (x_i - \bar{x})^2/(n-1) = a^2 s_x^2. \end{aligned} \tag{1.9}$$

With the above type of transformation, computations for $\bar{u}$ and s_u^2 become simple. Now, $\bar{x}$ is obtained from $(\bar{u} - b)/a$ and s_x^2 from s_u^2/a^2. Note that adding or subtracting a constant displaces the mean, but it has no effect on the variance. Multiplying x_i by the constant a results in multiplying its variance by a^2, and the standard deviation by a.

As found earlier, the average of the heights of the twenty boys is 175.2 cm. To convert x_i in cm to y_i in inches, $y_i = 0.3937\,x_i$. Now, the average height is $\bar{y} = (0.3937)\bar{x} = (0.3937)(175.2) = 68.98$ inches or close to 5 feet 9 inches. The variance becomes $s_y^2 = (0.3937)^2 s_x^2 = (0.3937)^2(51.33) = 7.9561$ and $s_y = (0,3937)s_x = 2.82$ inches.

1.10 Linear combinations of two sets of observations

Consider the gains in weights (x_i, y_i), $i = (1,2,\ldots,n)$ of a sample of n adults on two occasions. The total $t_i = (x_i + y_i)$, difference $d_i = (y_i - x_i)$, a weighted combination $u_i = (ax_i + by_i)$, with specified constants (a, b) may be of interest. The mean and variance of t_i are

$$\bar{t} = \sum_{1}^{n} t_i/n = \sum_{1}^{n} (x_i + y_i)/n = \bar{x} + \bar{y} \tag{1.10}$$

and

$$\mathrm{V}(t_\mathrm{i}) = \sum_{1}^{n} (t_i - \bar{t})^2/(n-1) = \sum_{1}^{n} \left[(x_i - \bar{x})^2 + (y_i - \bar{y})^2 + 2(x_i - \bar{x})(y_i - \bar{y}) \right] /(n-1)$$

$$= s_x^2 + s_y^2 + 2s_{xy}, \tag{1.11}$$

where $\left(s_x^2, s_y^2\right)$ are the variances and s_{xy} the covariance of x and y. The standard deviations of x and y are s_x and s_y, and the sample correlation is $r = s_\mathrm{xy}/s_\mathrm{x}s_\mathrm{y}$, $(-1 < r < 1)$. The standard deviation s_t of t_i is the positive square root of $V(t_i)$.

Similarly, the mean, variance, and standard deviation of d_i are $\bar{d} = \bar{y} - \bar{x}$, $V(d_i) = s_y^2 + s_x^2 - 2s_{xy}$ and the standard deviation s_d of d_i is the positive square root of $V(d_i)$. If $u_i = ax_i + by_i + c$, where (a, b, c) are constants, $\bar{u} = a\bar{x} + b\bar{y} + c$ and $V(u_i) = V(ax_i + by_i) = a^2 s_x^2 + b^2 s_y^2 + 2abs_{xy}$. The standard deviation s_u of u_i is obtained from the square root of $V(u_i)$.

For an illustration, consider the gains in weights (lbs) (x_i, y_i) of $n = 5$ adults on two occasions: (5, 10), (10, 5), (10, 10), (5, –5), and (5, 10); the fourth candidate lost 5 lbs.

From these observations, the mean, variance and standard deviation of x_i are (7, 7.5, 2.74). Corresponding figures for y_i are (6, 42.5, 6.52). The covariance and correlation are $s_{xy} = 3.75$ and $r = 0.21$. The mean, variance and standard deviation of t_i and d_i respectively become (13, 57.5, 7.58) and (–1, 42.5, 6.52). With $a = (1/4)$, $b = (3/4)$ and $c = -5$, the mean, variance, and standard deviation of u_i become (1.25, 57.78, 5.08).

Exercises

1.1. Find the summary figures for the 20 ten-year old boys and girls in Tables T1.1 and T1.2.

1.2. (a) Find the means and standard deviations of the three characteristics for the sixteen-year-old boys and girls in Tables T1.3 and T1.4. (b) Find the means and S.D.s for the heights with grouping.

1.3. The mid-values of weights (lbs.) along with the systolic blood pressures, SBPs, of 200 adults are presented below. Find the means and standard deviations of the weights and blood pressures and their correlation.

		SBP (y)		
		130	150	Total
Weight (x)	145	32	44	76
	155	28	40	68
	165	20	36	56
	Total	80	120	200

1.4. Fertility rates per woman (x in %) and the corresponding annual population growth rates (y in %) in 192 countries of the world in 2006 are available from the tables of the WHO (2008). The fertility rates ranged from 1.2 to 7.3 percent. The population growth rate was negative in 18 countries and ranged from 0 to 4 percent in 188 countries. Combining the very small values of (x, y) with the adjacent cells, the mid-values of the percentages and the frequencies are presented below. Find the means, standard deviations and the correlation of these two characteristics.

		Population growth rate (y %)				
		0.5	1.5	2.5	3.5	Total
	2	71	40	2	2	115
Fertility rate (x %)	3.5	8	13	8	1	30
	5.5	1	8	26	12	47
	Total	80	61	36	15	192

1.5. Ross et al. (2006) analyzed the use of healthcare services by the lower- and higher-income insured and uninsured adults in the United States. The data were obtained from a nationally representative survey of a sample of 194,943 adults conducted by the CDC in 2002. The responding size (n) of the sample and the percentages of the insured (I) and uninsured (U) for the age, income and household classifications were presented as follows. Estimate the means, medians and standard deviations of age, income and household size for the insured and uninsured.

Age	I	U	Income ($ 1,000)	I	U	Household size	I	U
18–29	24	38	< 15	8	23	1	11	10
30–39	24	24	15–25	12	34	2	29	24
40–49	25	21	25–35	13	18	3	21	21
50–64	27	18	35–50	19	13	4	22	20
			50-75	21	7	5	17	25
			≥ 75	27	5			
$n = 194,943$			$n = 172,778$			$n = 194,695$		

1.6. Immunization coverage of the one-year-olds in the countries of the world for measles, DTP3 and HepB3 are presented in Table T1.5. Find the means and standard deviations for each of these types of coverage.

1.7. Convert the average and standard deviation of the weights in Table 1.2 into pounds from kilograms.

1.8. With the observations in Section (1.10), find the means, variances and standard deviations of (a) $u_i = (1/2)x_i + (1/2)y_i - 10$ and (b) $v_i = (1/3)x_i + (2/3)y_i + 5$.

1.9. Find the covariance and correlation of u_i and v_i of Exercise 1.8.