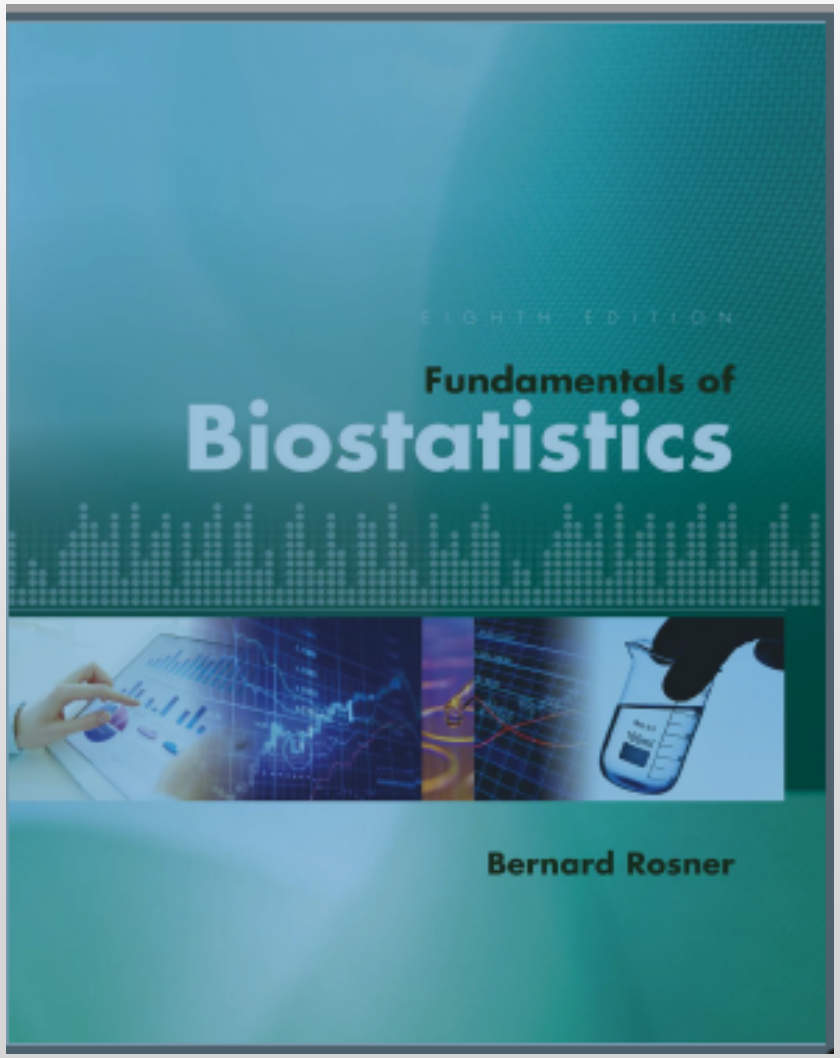
Biostatistic


EIGHTH EDITION
Fundamentals of
Biostatistics
Bernard Rosner

Statistics

whereby: vasıtasıyla
inferences: çıkarım

- **Statistics** is the science whereby inferences are made about specific random phenomena on the basis of relatively limited sample material.
- The field of statistics has two main areas:
  - mathematical statistics
  - applied statistics.

Statistics

inference: çıkarım
implementation: uygulama

- **Mathematical statistics** concerns the development of new methods of statistical inference and requires detailed knowledge of abstract mathematics for its implementation.

- **Applied statistics** involves applying the methods of mathematical statistics to specific subject areas, such as economics, psychology, and public health.

Biostatistics

- Biostatistics is the branch of applied statistics that applies statistical methods to medical and biological problems.

Descriptive Statistics

- The Arithmetic Mean
- The Median
- The Mode

Colloquially: halk dilinde

The Arithmetic Mean

- One measure of location for any sample is the arithmetic mean (colloquially called the *average*).

- The arithmetic mean (or mean or sample mean) is usually denoted by x.

DEFINITION 2.1 The **arithmetic mean** is the sum of all the observations divided by the number of observations. It is written in statistical terms as

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The sign Σ (sigma) in Definition 2.1 is a summation sign. The expression

$$\sum_{i=1}^{n} x_i$$

is simply a short way of writing the quantity $(x_1 + x_2 + \cdots + x_n)$.

If a and b are integers, where $a < b$, then

$$\sum_{i=a}^{b} x_i$$

means $x_a + x_{a+1} + \cdots + x_b$.

If $a = b$, then $\sum_{i=a}^{b} x_i = x_a$. One property of summation signs is that if each term in the summation is a multiple of the same constant c, then c can be factored out from the summation; that is,

$$\sum_{i=1}^{n} cx_i = c\left(\sum_{i=1}^{n} x_i\right)$$

EXAMPLE 2.3

If $x_1 = 2 \quad x_2 = 5 \quad x_3 = -4$

find $\sum_{i=1}^{3} x_i \quad \sum_{i=2}^{3} x_i \quad \sum_{i=1}^{3} x_i^2 \quad \sum_{i=1}^{3} 2x_i$

Solution:

$$\sum_{i=1}^{3} x_i = 2+5-4=3 \quad \sum_{i=2}^{3} x_i = 5-4=1$$

$$\sum_{i=1}^{3} x_i^2 = 4+25+16=45 \quad \sum_{i=1}^{3} 2x_i = 2\sum_{i=1}^{3} x_i = 6$$

The Arithmetic Mean

- How to define the middle of a sample may seem obvious,
- but the more you think about it, the less obvious it becomes.
- Suppose the sample consists of the birthweights of all live-born infants born at a hospital in San Diego, California, during a 1-week period.
- This sample is shown in Table 2.1.

TABLE 2.1 Sample of birthweights (g) of live-born infants born at a private hospital in San Diego, California, during a 1-week period

i	x_i	i	x_i	i	x_i	i	x_i
1	3265	6	3323	11	2581	16	2759
2	3260	7	3649	12	2841	17	3248
3	3245	8	3200	13	3609	18	3314
4	3484	9	3031	14	2838	19	3101
5	4146	10	2069	15	3541	20	2834

EXAMPLE 2.4 What is the arithmetic mean for the sample of birthweights in Table 2.1?

$$\bar{x} = (3265 + 3260 + \cdots + 2834)/20 = 3166.9 \text{ g}$$

TABLE 2.1 Sample of birthweights (g) of live-born infants born at a private hospital in San Diego, California, during a 1-week period

i	x_i	i	x_i	i	x_i	i	x_i
1	3265	6	3323	11	2581	16	2759
2	3260	7	3649	12	2841	17	3248
3	3245	8	3200	13	3609	18	3314
4	3484	9	3031	14	2838	19	3101
5	4146	10	2069	15	3541	20	2834

what is the age arithmetic mean of our class ??

TABLE 2.1 Sample of birthweights (g) of live-born infants born at a private hospital in San Diego, California, during a 1-week period

i	x_i	i	x_i	i	x_i	i	x_i
1	3265	6	3323	11	2581	16	2759
2	3260	7	3649	12	2841	17	3248
3	3245	8	3200	13	3609	18	3314
4	3484	9	3031	14	2838	19	3101
5	4146	10	2069	15	3541	20	2834

- The arithmetic mean is, in general, a very natural measure of location.
- One of its main limitations, however, is that it is oversensitive to extreme values.
- In this instance, it may not be representative of the location of the great majority of sample points.
- For example, if the first infant in Table 2.1 happened to be a premature infant weighing 500 g rather than 3265 g, then the arithmetic mean of the sample would fall to 3028.7 g.
- In this instance, 6 of the birthweights would be lower than the arithmetic mean, and 14 would be higher than the arithmetic mean.

TABLE 2.1 Sample of birthweights (g) of live-born infants born at a private hospital in San Diego, California, during a 1-week period

i	x_i	i	x_i	i	x_i	i	x_i
1	3265	6	3323	11	2581	16	2759
2	3260	7	3649	12	2841	17	3248
3	3245	8	3200	13	3609	18	3314
4	3484	9	3031	14	2838	19	3101
5	4146	10	2069	15	3541	20	2834

- It is possible in extreme cases for all but one of the sample points to be on one side of the arithmetic mean.
- In these types of samples, the arithmetic mean is a poor measure of central location because it does not reflect the center of the sample.
- Nevertheless, the arithmetic mean is by far the most widely used measure of central location.

The Median

order: sıraya koymak
Odd: tek
Even:çift

- An alternative measure of location, perhaps second in popularity to the arithmetic mean, is the median.

- Suppose there are *n* observations in a sample.
- If these observations are ordered from **smallest to largest**, then the median is defined as follows:

DEFINITION 2.2 The sample median is

(1) The $\left(\frac{n+1}{2}\right)$th largest observation if n is odd

(2) The average of the $\left(\frac{n}{2}\right)$th and $\left(\frac{n}{2}+1\right)$th largest observations if n is even

- The rationale for these definitions is to ensure an equal number of sample points on both sides of the sample median. (Bu tanımların mantığı,örnek medyanının her iki tarafında eşit sayıda numune noktası sağlamaktır.)

- The median is defined differently when *n* is even and odd because it is impossible to achieve this goal with one uniform definition.
- Samples with an odd sample size have a unique central point;
- for example, for samples of size 7, the fourth largest point is the central point in the sense that 3 points are smaller than it and 3 points are larger

- Samples with an even sample size have no unique central point, and the middle two values must be averaged.
- Thus, for samples of size 8 the fourth and fifth largest points would be averaged to obtain the median, because neither is the central point.

EXAMPLE 2.5

Compute the sample median for the sample in Table 2.1.

Solution: First, arrange the sample in ascending order:

2069, 2581, 2759, 2834, 2838, 2841, 3031, 3101, 3200, 3245, 3248, 3260, 3265, 3314, 3323, 3484, 3541, 3609, 3649, 4146

Because n is even,

Sample median = average of the 10th and 11th largest observations
= (3245 + 3248)/2 = 3246.5 g

TABLE 2.1 Sample of birthweights (g) of live-born infants born at a private hospital in San Diego, California, during a 1-week period

i	x_i	i	x_i	i	x_i	i	x_i
1	3265	6	3323	11	2581	16	2759
2	3260	7	3649	12	2841	17	3248
3	3245	8	3200	13	3609	18	3314
4	3484	9	3031	14	2838	19	3101
5	4146	10	2069	15	3541	20	2834

EXAMPLE 2.6 **Infectious Disease** Consider the data set in Table 2.3, which consists of white-blood counts taken upon admission of all patients entering a small hospital in Allentown, Pennsylvania, on a given day. Compute the median white-blood count.

TABLE 2.3 **Sample of admission white-blood counts (× 1000) for all patients entering a hospital in Allentown, Pennsylvania, on a given day**

i	x_i	i	x_i
1	7	6	3
2	35	7	10
3	5	8	12
4	9	9	8
5	8		

Solution: First, order the sample as follows: 3, 5, 7, 8, 8, 9, 10, 12, 35. Because n is odd, the sample median is given by the fifth largest point, which equals 8 or 8000 on the original scale.

TABLE 2.3 Sample of admission white-blood counts (× 1000) for all patients entering a hospital in Allentown, Pennsylvania, on a given day

i	x_i	i	x_i
1	7	6	3
2	35	7	10
3	5	8	12
4	9	9	8
5	8		

- The main strength of the sample median is that it is insensitive to very large or very small values.
- In particular, if the second patient in Table 2.3 had a white count of 65,000 rather than 35,000, the sample median would remain unchanged, because the fifth largest value is still 8000.
- Conversely, the arithmetic mean would increase dramatically from 10,778 in the original sample to 14,111 in the new sample.
- The main weakness of the sample median is that it is determined mainly by the middle points in a sample and is less sensitive to the actual numeric values of the remaining data points.

The Mode

- Another widely used measure of location is the mode.

DEFINITION 2.3 The **mode** is the most frequently occurring value among all the observations in a sample.

EXAMPLE 2.7

Gynecology Consider the sample of time intervals between successive menstrual periods for a group of 500 college women age 18 to 21 years, shown in Table 2.4. The frequency column gives the number of women who reported each of the respective durations. The mode is 28 because it is the most frequently occurring value.

TABLE 2.4 Sample of time intervals between successive menstrual periods (days) in college-age women

Value	Frequency	Value	Frequency	Value	Frequency
24	5	29	96	34	7
25	10	30	63	35	3
26	28	31	24	36	2
27	64	32	9	37	1
28	185	33	2	38	1

EXAMPLE 2.8 Compute the mode of the distribution in Table 2.3.

Solution: The mode is 8000 because it occurs more frequently than any other white-blood count.

TABLE 2.3 Sample of admission white-blood counts ($\times$ 1000) for all patients entering a hospital in Allentown, Pennsylvania, on a given day

i	x_i	i	x_i
1	7	6	3
2	35	7	10
3	5	8	12
4	9	9	8
5	8		

- Some distributions have more than one mode.
- In fact, one useful method of classifying distributions is by the number of modes present.
- A distribution with
  - one mode is called unimodal;
  - two modes, bimodal;
  - three modes, trimodal;
  - and so forth.

EXAMPLE 2.9 Compute the mode of the distribution in Table 2.1.

Solution: There is no mode, because all the values occur exactly once.

TABLE 2.1 Sample of birthweights (g) of live-born infants born at a private hospital in San Diego, California, during a 1-week period

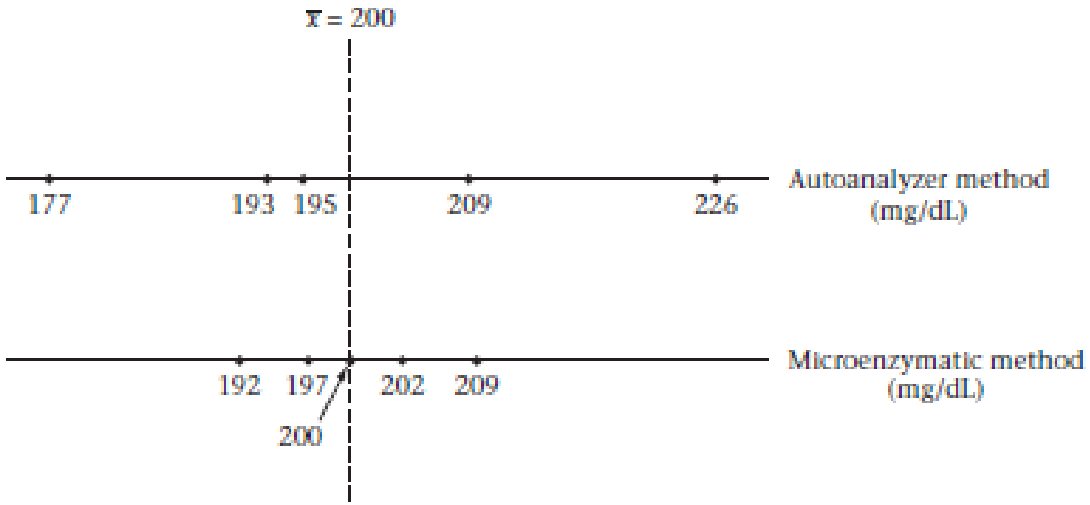
i	x_i	i	x_i	i	x_i	i	x_i
1	3265	6	3323	11	2581	16	2759
2	3260	7	3649	12	2841	17	3248
3	3245	8	3200	13	3609	18	3314
4	3484	9	3031	14	2838	19	3101
5	4146	10	2069	15	3541	20	2834

- Example 2.9 illustrates a common problem with the mode:
- It is not a useful measure of location if there is a large number of possible values, each of which occurs infrequently.
- In such cases the mode will be either far from the center of the sample or, in extreme cases, will not exist, as in Example 2.9.

Measures of Spread

- Consider Figure 2.4, which represents two samples of cholesterol measurements, each on the same person, but using different measurement techniques. (attention)
- The samples appear to have about the same center, and whatever measure of central location is used is probably about the same in the two samples.
- In fact, the arithmetic means are both 200 mg/dL.

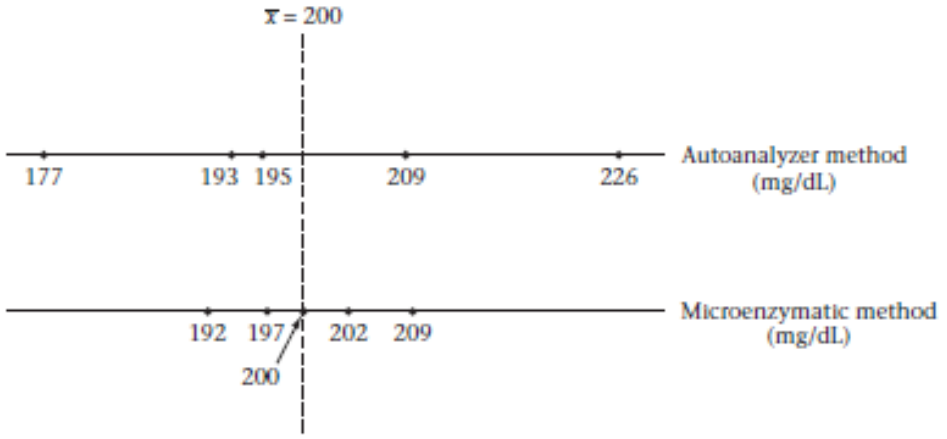
FIGURE 2.4 Two samples of cholesterol measurements on a given person using the Autoanalyzer and Microenzymatic measurement methods



Measures of Spread

- Visually, however, the two samples appear radically different.
- This difference lies in the greater variability, or spread, of the Autoanalyzer method relative to the Microenzymatic method.
- In this section, the notion of variability is quantified.
- Many samples can be well described by a combination of a measure of location and a measure of spread.

FIGURE 2.4 Two samples of cholesterol measurements on a given person using the Autoanalyzer and Microenzymatic measurement methods



The Range

- Several different measures can be used to describe the variability of a sample.
- Perhaps the simplest measure is the range.

DEFINITION 2.5 The **range** is the difference between the largest and smallest observations in a sample.

Range=Max-min

EXAMPLE 2.14 The range in the sample of birthweights in Table 2.1 is

$$4146 - 2069 = 2077 \text{ g}$$

TABLE 2.1 Sample of birthweights (g) of live-born infants born at a private hospital in San Diego, California, during a 1-week period

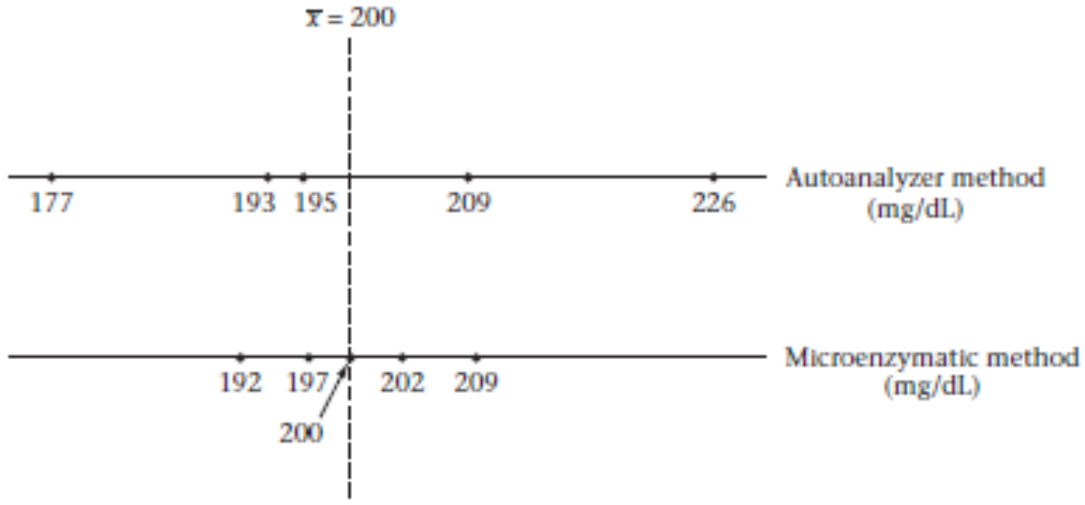
i	x_i	i	x_i	i	x_i	i	x_i
1	3265	6	3323	11	2581	16	2759
2	3260	7	3649	12	2841	17	3248
3	3245	8	3200	13	3609	18	3314
4	3484	9	3031	14	2838	19	3101
5	4146	10	2069	15	3541	20	2834

EXAMPLE 2.15

Compute the ranges for the Autoanalyzer- and Microenzymatic-method data in Figure 2.4, and compare the variability of the two methods.

Solution: The range for the Autoanalyzer method = 226 – 177 = 49 mg/dL. The range for the Microenzymatic method = 209 – 192 = 17 mg/dL. The Autoanalyzer method clearly seems more variable.

FIGURE 2.4 Two samples of cholesterol measurements on a given person using the Autoanalyzer and Microenzymatic measurement methods



- One advantage of the range is that it is very easy to compute once the sample points are ordered.
- One striking disadvantage is that it is very sensitive to extreme observations.
- Hence, if the lightest infant in Table 2.1 weighed 500 g rather than 2069 g, then the range would increase dramatically to 4146 − 500 = 3646 g.

TABLE 2.1 Sample of birthweights (g) of live-born infants born at a private hospital in San Diego, California, during a 1-week period

i	x_i	i	x_i	i	x_i	i	x_i
1	3265	6	3323	11	2581	16	2759
2	3260	7	3649	12	2841	17	3248
3	3245	8	3200	13	3609	18	3314
4	3484	9	3031	14	2838	19	3101
5	4146	10	2069	15	3541	20	2834

- Another disadvantage of the range is that it depends on the sample size (*n*).
- That is, the larger sample size (*n*) is, the larger the range tends to be.
- This complication makes it difficult to compare ranges from data sets of differing size.

Quantiles (Percentiles)

Shortcomings: eksiklikler

- Another approach that addresses some of the shortcomings of the range in quantifying the spread in a data set is the use of quantiles or percentiles.
- A procedure for obtaining the p^{th} percentile of a data set of size n is as follows:
- Step1: Arrange the data in ascending (increasing) order
- Step 2: Compute an index i as follows $i=p.n/100$
- Step3:
  - If i is an integer, the p^{th} percentile is the average of the i^{th} and $(i+1)^{th}$ smallest data values
  - If i is not an integer then round i up to the nearest integer and take the value at that postion

Quantiles (Percentiles)

- The median, being the 50th percentile, is a special case of a quantile.

DEFINITION 2.6 The ***p*th percentile** is defined by

(1) The $(k + 1)$th largest sample point if $np/100$ is not an integer (where k is the largest integer less than $np/100$).

(2) The average of the $(np/100)$th and $(np/100 + 1)$th largest observations if $np/100$ is an integer.

Percentiles are also sometimes called **quantiles**.

- The spread of a distribution can be characterized by specifying several percentiles.
- For example, the 10th and 90th percentiles are often used to characterize spread.
- Percentiles have the advantage over the range of being less sensitive to outliers and of not being greatly affected by the sample size (*n*).

EXAMPLE 2.16

Compute the 10th and 90th percentiles for the birthweight data in Table 2.1.

Solution: Because $20 \times .1 = 2$ and $20 \times .9 = 18$ are integers, the 10th and 90th percentiles are defined by

10th percentile: average of the second and third largest values
$= (2581 + 2759)/2 = 2670$ g

90th percentile: average of the 18th and 19th largest values
$= (3609 + 3649)/2 = 3629$ g

We would estimate that 80% of birthweights will fall between 2670 g and 3629 g, which gives an overall impression of the spread of the distribution.

TABLE 2.1 Sample of birthweights (g) of live-born infants born at a private hospital in San Diego, California, during a 1-week period

i	x_i	i	x_i	i	x_i	i	x_i
1	3265	6	3323	11	2581	16	2759
2	3260	7	3649	12	2841	17	3248
3	3245	8	3200	13	3609	18	3314
4	3484	9	3031	14	2838	19	3101
5	4146	10	2069	15	3541	20	2834

2069, 2581, 2759, 2834, 2838, 2841, 3031, 3101, 3200, 3245, 3248, 3260, 3265, 3314, 3323, 3484, 3541, 3609, 3649, 4146

do not forget.

EXAMPLE 2.17 **Compute the 20th percentile for the white-blood-count data in Table 2.3.**

Solution: Because $np/100 = 9 \times .2 = 1.8$ is not an integer, the 20th percentile is defined by the (1 + 1)th largest value = second largest value = 5000.

TABLE 2.3 Sample of admission white-blood counts (× 1000) for all patients entering a hospital in Allentown, Pennsylvania, on a given day

i	x_i	i	x_i
1	7	6	3
2	35	7	10
3	5	8	12
4	9	9	8
5	8		

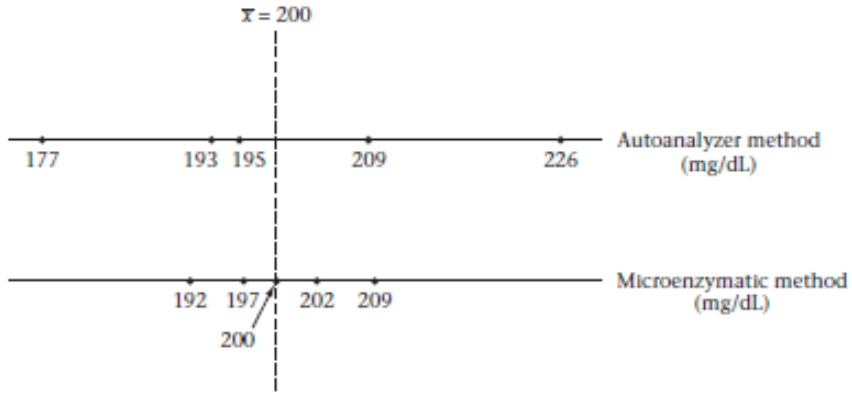
Percentiles not only give locate the center of a distribution but also other locations in a distribution

our class ? Percentiles

The Variance and Standard Deviation

- The main difference between the Autoanalyzer- and Microenzymatic-method data in Figure 2.4 is that the Microenzymatic-method values are closer to the center of the sample than the Autoanalyzer-method values.
- If the center of the sample is defined as the arithmetic mean, then a measure that can summarize the difference (or deviations) between the individual sample points and the arithmetic mean is needed.

FIGURE 2.4 Two samples of cholesterol measurements on a given person using the Autoanalyzer and Microenzymatic measurement methods



The Variance and Standard Deviation Example

- Biomedical Engineering age
- Electronics engineer ig age

Variance and Standard Deviation

- The most important measures of dispersion are the variance and its square root, the standard deviation.
- Since the variance is just the square of the standard deviation, these quantities contain essentially the same information, just on different scales.

DEFINITION 2.7 The **sample variance**, or **variance**, is defined as follows:

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

DEFINITION 2.8 The **sample standard deviation**, or **standard deviation**, is defined as follows:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\text{sample variance}}$$

EXAMPLE 2.19 Compute the variance and standard deviation for the Autoanalyzer- and Microenzymatic-method data in Figure 2.4.

Solution: Autoanalyzer Method

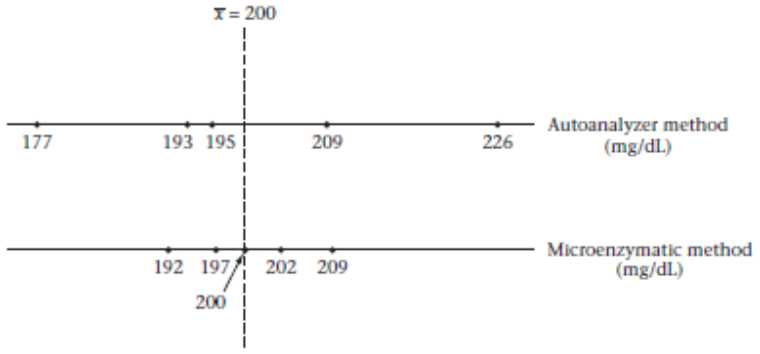
$$s^2 = \left[(177-200)^2 + (193-200)^2 + (195-200)^2 + (209-200)^2 + (226-200)^2\right]/4$$
$$= (529+49+25+81+676)/4 = 1360/4 = 340$$
$$s = \sqrt{340} = 18.4$$

Microenzymatic Method

$$s^2 = \left[(192-200)^2 + (197-200)^2 + (200-200)^2 + (202-200)^2 + (209-200)^2\right]/4$$
$$= (64+9+0+4+81)/4 = 158/4 = 39.5$$
$$s = \sqrt{39.5} = 6.3$$

Thus the Autoanalyzer method has a standard deviation roughly three times as large as that of the Microenzymatic method.

FIGURE 2.4 Two samples of cholesterol measurements on a given person using the Autoanalyzer and Microenzymatic measurement methods



EXAMPLE 2.20

Use Microsoft Excel to compute the mean and standard deviation for the Autoanalyzer and Microenzymatic-method data in Figure 2.4.

Solution: We enter the Autoanalyzer and Microenzymatic data in cells B3–B7 and C3–C7, respectively. We then use the Average and StDev functions to evaluate the mean and standard deviation as follows:

	Autoanalyzer Method	Microenzymatic Method
	177	192
	193	197
	195	200
	209	202
	226	209
Average	200	200
StDev	18.4	6.3

In Excel, if we make B8 the active cell and type = Average(B3:B7) in that cell, then the mean of the values in cells B3, B4, . . . , B7 will appear in cell B8. Similarly, specifying = Stdev(B3:B7) will result in the standard deviation of the Autoanalyzer Method data being placed in the active cell of the spreadsheet.

EXAMPLE 2.22 Compute the variance and standard deviation of the birthweight data in Table 2.1 in both grams and ounces.

Solution: The original data are given in grams, so first compute the variance and standard deviation in these units.

$$s^2 = \frac{(3265-3166.9)^2 + \cdots + (2834-3166.9)^2}{19}$$
$$= 3{,}768{,}147.8/19 = 198{,}323.6 \text{ g}^2$$
$$s = 445.3 \text{ g}$$

To compute the variance and standard deviation in ounces, note that

$$1 \text{ oz} = 28.35 \text{ g} \quad \text{or} \quad y_i = \frac{1}{28.35} x_i$$

Thus, $s^2(\text{oz}) = \frac{1}{28.35^2} s^2(\text{g}) = 246.8 \text{ oz}^2$

$$s(\text{oz}) = \frac{1}{28.35} s(\text{g}) = 15.7 \text{ oz}$$

The Coefficient of Variation

- It is useful to relate the arithmetic mean and the standard deviation to each other because, for example, a standard deviation of 10 means something different conceptually if the arithmetic mean is 10 versus if it is 1000.
- A special measure, the coefficient of variation, is often used for this purpose.

DEFINITION 2.9 The coefficient of variation (CV) is defined by

$$100\% \times (s/\overline{x})$$

- This measure remains the same regardless of what units are used
- because if the units change by a factor *c*,
- then both the mean and standard deviation change by the factor *c;*
- while the *CV*, which is the ratio between them, remains unchanged.

EXAMPLE 2.23

Compute the coefficient of variation for the data in Table 2.1 when the birthweights are expressed in either grams or ounces.

Solution: $CV = 100\% \times (s/\overline{x}) = 100\% \times (445.3\,\text{g}/3166.9\,\text{g}) = 14.1\%$

If the data were expressed in ounces, then

$$CV = 100\% \times (15.7\ \text{oz}/111.71\ \text{oz}) = 14.1\%$$

TABLE 2.1 Sample of birthweights (g) of live-born infants born at a private hospital in San Diego, California, during a 1-week period

i	x_i	i	x_i	i	x_i	i	x_i
1	3265	6	3323	11	2581	16	2759
2	3260	7	3649	12	2841	17	3248
3	3245	8	3200	13	3609	18	3314
4	3484	9	3031	14	2838	19	3101
5	4146	10	2069	15	3541	20	2834

- The *CV* is most useful in comparing the variability of several different samples, each with different arithmetic means.

- This is because a higher variability is usually expected when the mean increases, and the *CV* is a measure that accounts for this variability.

Grouped Data

- Sometimes the sample size is too large to display all the raw data.
- Also, data are frequently collected in grouped
- Although a set of observation can be made more comprehensible and meaningful by means of an ordered array, further useful summarization may be achieved by grouping the data.

Grouped Data

- Consider the data set in Table 2.9, which represents the birthweights from 100 consecutive deliveries at a Boston hospital.
- Suppose we wish to display these data for publication purposes.
- How can we do this?
- The simplest way to display the data is to generate a frequency distribution using a statistical package.

TABLE 2.9 Sample of birthweights (oz) from 100 consecutive deliveries at a Boston hospital

58	118	92	108	132	32	140	138	96	161
120	86	115	118	95	83	112	128	127	124
123	134	94	67	124	155	105	100	112	141
104	132	98	146	132	93	85	94	116	113
121	68	107	122	126	88	89	108	115	85
111	121	124	104	125	102	122	137	110	101
91	122	138	99	115	104	98	89	119	109
104	115	138	105	144	87	88	103	108	109
128	106	125	108	98	133	104	122	124	110
133	115	127	135	89	121	112	135	115	64

TABLE 2.10 Frequency distribution of the birthweight data on Table 2.9 using the FREQ procedure of SAS

Birthweight	Frequency	Percent	Cumulative Frequency	Cumulative Percent
32	1	1.00	1	1.00
58	1	1.00	2	2.00
64	1	1.00	3	3.00
67	1	1.00	4	4.00
68	1	1.00	5	5.00
83	1	1.00	6	6.00
85	2	2.00	8	8.00
86	1	1.00	9	9.00
87	1	1.00	10	10.00
88	2	2.00	12	12.00
89	3	3.00	15	15.00
91	1	1.00	16	16.00
92	1	1.00	17	17.00
93	1	1.00	18	18.00
94	2	2.00	20	20.00
95	1	1.00	21	21.00
96	1	1.00	22	22.00
98	3	3.00	25	25.00
99	1	1.00	26	26.00
100	1	1.00	27	27.00
101	1	1.00	28	28.00
102	1	1.00	29	29.00
103	1	1.00	30	30.00
104	5	5.00	35	35.00
105	2	2.00	37	37.00
106	1	1.00	38	38.00
107	1	1.00	39	39.00
108	4	4.00	43	43.00
109	2	2.00	45	45.00
110	2	2.00	47	47.00
111	1	1.00	48	48.00
112	3	3.00	51	51.00
113	1	1.00	52	52.00
115	6	6.00	58	58.00
116	1	1.00	59	59.00
118	2	2.00	61	61.00
119	1	1.00	62	62.00
120	1	1.00	63	63.00
121	3	3.00	66	66.00
122	4	4.00	70	70.00
123	1	1.00	71	71.00
124	4	4.00	75	75.00
125	2	2.00	77	77.00
126	1	1.00	78	78.00
127	2	2.00	80	80.00
128	2	2.00	82	82.00
132	3	3.00	85	85.00
133	2	2.00	87	87.00
134	1	1.00	88	88.00
135	2	2.00	90	90.00
137	1	1.00	91	91.00
138	3	3.00	94	94.00
140	1	1.00	95	95.00
141	1	1.00	96	96.00
144	1	1.00	97	97.00
146	1	1.00	98	98.00
155	1	1.00	99	99.00
161	1	1.00	100	100.00

DEFINITION 2.10

A **frequency distribution** is an ordered display of each value in a data set together with its **frequency**, that is, the number of times that value occurs in the data set. In addition, the percentage of sample points that take on a particular value is also typically given.

- A frequency distribution of the sample of 100 birthweights in Table 2.9, generated is displayed in Table 2.10.
- frequency procedure provides the
  - Frequency,
  - relative frequency (Percent),
  - Cumulative Frequency,
  - Cumulative Percent for each birthweight present in the sample.

- For any particular birthweight *b*,
- The Cumulative Frequency is the number of birthweights in the sample that are less than or equal to *b.*
- The Percent = 100 × Frequency/*n*,
- Cumulative Percent = 100 × Cumulative Frequency/*n*
- Cumulative Percent = the percentage of birthweights less than or equal to *b.*

TABLE 2.10 Frequency distribution of the birthweight data on Table 2.9 using the FREQ procedure of SAS

Birthweight	Frequency	Percent	Cumulative Frequency	Cumulative Percent
32	1	1.00	1	1.00
58	1	1.00	2	2.00
64	1	1.00	3	3.00
67	1	1.00	4	4.00
68	1	1.00	5	5.00
83	1	1.00	6	6.00
85	2	2.00	8	8.00
86	1	1.00	9	9.00
87	1	1.00	10	10.00
88	2	2.00	12	12.00
89	3	3.00	15	15.00
91	1	1.00	16	16.00
92	1	1.00	17	17.00

- 1 oz =28.35 gram

TABLE 2.9 Sample of birthweights (oz) from 100 consecutive deliveries at a Boston hospital

58	118	92	108	132	32	140	138	96	161
120	86	115	118	95	83	112	128	127	124
123	134	94	67	124	155	105	100	112	141
104	132	98	146	132	93	85	94	116	113
121	68	107	122	126	88	89	108	115	85
111	121	124	104	125	102	122	137	110	101
91	122	138	99	115	104	98	89	119	109
104	115	138	105	144	87	88	103	108	109
128	106	125	108	98	133	104	122	124	110
133	115	127	135	89	121	112	135	115	64

TABLE 2.10 **Frequency distribution of the birthweight data on Table 2.9 using the FREQ procedure of SAS**

Birthweight	Frequency	Percent	Cumulative Frequency	Cumulative Percent
32	1	1.00	1	1.00
58	1	1.00	2	2.00
64	1	1.00	3	3.00
67	1	1.00	4	4.00
68	1	1.00	5	5.00
83	1	1.00	6	6.00
85	2	2.00	8	8.00
86	1	1.00	9	9.00
87	1	1.00	10	10.00
88	2	2.00	12	12.00
89	3	3.00	15	15.00
91	1	1.00	16	16.00
92	1	1.00	17	17.00
93	1	1.00	18	18.00
94	2	2.00	20	20.00
95	1	1.00	21	21.00
96	1	1.00	22	22.00
98	3	3.00	25	25.00
99	1	1.00	26	26.00
100	1	1.00	27	27.00
101	1	1.00	28	28.00
102	1	1.00	29	29.00
103	1	1.00	30	30.00
104	5	5.00	35	35.00
105	2	2.00	37	37.00
106	1	1.00	38	38.00
107	1	1.00	39	39.00
108	4	4.00	43	43.00
109	2	2.00	45	45.00
110	2	2.00	47	47.00
111	1	1.00	48	48.00
112	3	3.00	51	51.00
113	1	1.00	52	52.00
115	6	6.00	58	58.00
116	1	1.00	59	59.00
118	2	2.00	61	61.00
119	1	1.00	62	62.00
120	1	1.00	63	63.00
121	3	3.00	66	66.00
122	4	4.00	70	70.00
123	1	1.00	71	71.00
124	4	4.00	75	75.00
125	2	2.00	77	77.00
126	1	1.00	78	78.00
127	2	2.00	80	80.00
128	2	2.00	82	82.00
132	3	3.00	85	85.00
133	2	2.00	87	87.00
134	1	1.00	88	88.00
135	2	2.00	90	90.00
137	1	1.00	91	91.00
138	3	3.00	94	94.00
140	1	1.00	95	95.00
141	1	1.00	96	96.00
144	1	1.00	97	97.00
146	1	1.00	98	98.00
155	1	1.00	99	99.00
161	1	1.00	100	100.00

- The main purpose of grouping is to make the data better understood and look better.
- The people who collect the data can make groupings as they wish.
- They can divide them into as many groups as they want.
- Make sure that the data is understandable when dividing into groups.
- When grouping, the amount of data is very important.
- An example of grouping by data numbers.
- But you do not have to do so.
- you can do as you wish.

if there is 20 data, make 5 groups
if there is 50 data, make 7 groups
if there is 100 data, make 8 groups
if there is 1000 data, make 11 groups

TABLE 2.11 General layout of grouped data

Group interval	Frequency
$y_1 \le x < y_2$	f_1
$y_2 \le x < y_3$	f_2
.	.
.	.
.	.
$y_i \le x < y_{i+1}$	f_i
.	.
.	.
.	.
$y_k \le x < y_{k+1}$	f_k

TABLE 2.12 Grouped frequency distribution of the birthweight (oz) from 100 consecutive deliveries

The FREQ Procedure

Group_interval	Frequency	Percent	Cumulative Frequency	Cumulative Percent
$29.5 \le x < 69.5$	5	5.00	5	5.00
$69.5 \le x < 89.5$	10	10.00	15	15.00
$89.5 \le x < 99.5$	11	11.00	26	26.00
$99.5 \le x < 109.5$	19	19.00	45	45.00
$109.5 \le x < 119.5$	17	17.00	62	62.00
$119.5 \le x < 129.5$	20	20.00	82	82.00
$129.5 \le x < 139.5$	12	12.00	94	94.00
$139.5 \le x < 169.5$	6	6.00	100	100.00