# Chapter 02

## Descriptive Statistics

# Introduction

The first step in data analysis is to describe the data in some concise manner.

Descriptive statistics that involve numeric or graphic display are crucial in capturing and conveying the final results of studies in publications.

Features of good numeric or graphic form of data summarization:

- Self-contained
- Understandable without reading the text
- Clearly labeled of attributes with well-defined terms
- Indicate principal trends in data

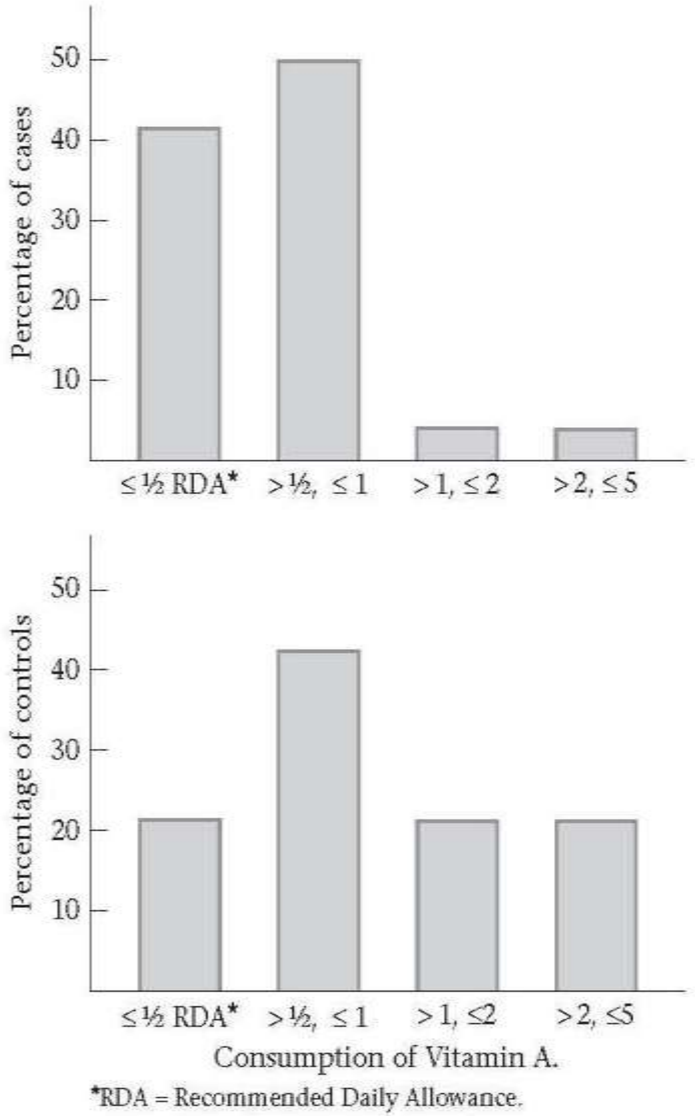
# ns

n A consumption
s cancer

ncer cases: 200
atched controls: 200

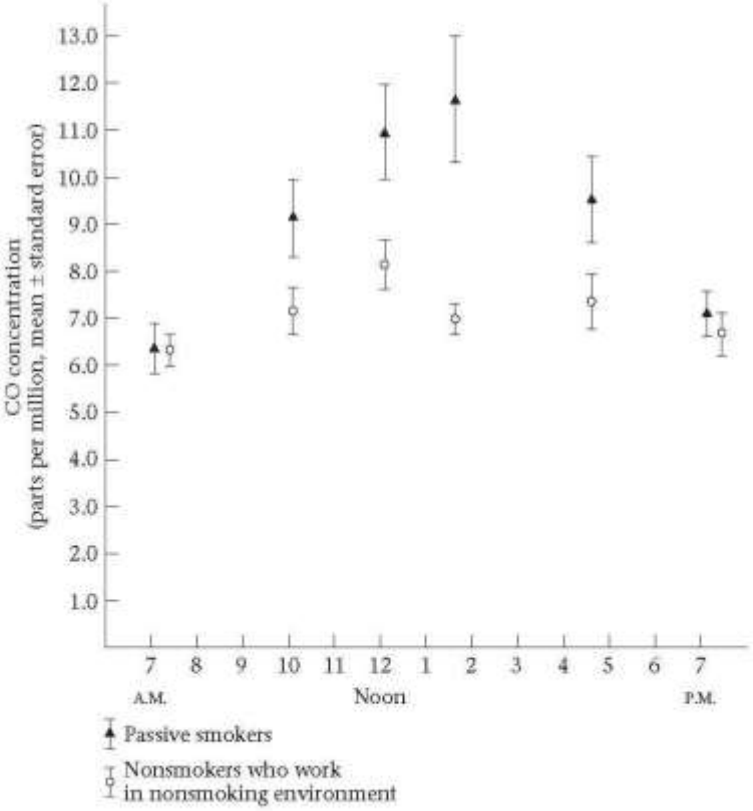
graphs show that the
A consumed by
is more than that
ed by the patients with
In some cases, the
exceed the
ended daily allowance

**Figure 2.1** Daily vitamin-A consumption among cancer cases and controls



# Example: Scatter plot

**Figure 2.2** Mean carbon-monoxide concentration (± standard error) by time of day as measured in the working environment of passive smokers and in nonsmokers who work in a nonsmoking environment



*Source:* Reproduced with permission of *The New England Journal of Medicine,* 302, 720–723, 1980.

CO concentrations are about the same in the working environments of passive smokers and non
smokers early in the day.

This supports the observation that passive smokers have lower pulmonary function than comparable nonsmokers.

# Measures of Location

It is easy to lose track of the overall picture when there are too many sample points.

Data summarization is important before any inferences can be made about the population from which the sample points have been obtained.

**Measure of location** is a type of measure useful for data summarization that defines the center or middle of the sample.

# Defining the Middle: The Arithmetic Mean

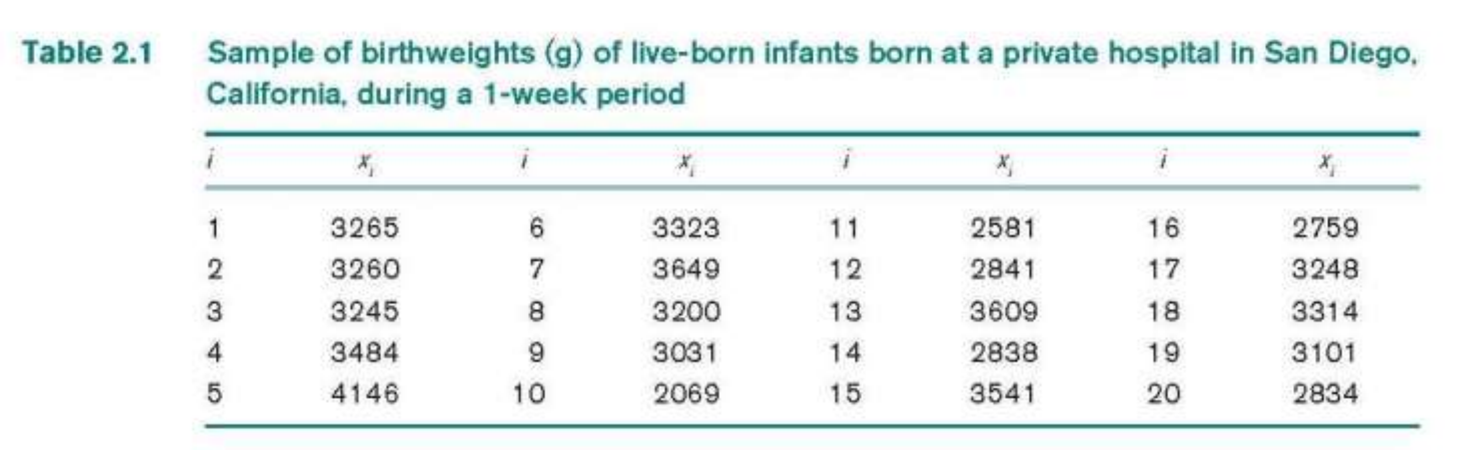


**Table 2.1** Sample of birthweights (g) of live-born infants born at a private hospital in San Diego, California, during a 1-week period

| $i$ | $x_i$ | $i$ | $x_i$ | $i$ | $x_i$ | $i$ | $x_i$ |
|---|---|---|---|---|---|---|---|
| 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 |

Arithmetic mean: the sum of all the observations divided by the number of observations.

Statistically expressed as

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

Limitation: Oversensitive to extreme values; in which case, it may not be representative of the location of the majority of sample points.

# Arithmetic Mean Explained

Si $\sum_{i=1}^{n} x_i$ summation sign.

lies $(x_1 + x_2 + \ldots + x_n)$

- If a and b are integers where a ≤ b, then $\sum_{i=a}^{b} x_i$ ning $(x_a + x_{a+1} + \ldots + x_b)$

- If a = b, then $\sum_{i=a}^{b} x_i = x_a$

- If c is some constant, then $\sum_{i=1}^{n} cx_i = c\left(\sum_{i=1}^{n} x_i\right)$

# Median

Sample median is

- $\left(\frac{n+1}{2}\right)$th the largest observation if n is odd
- Average of the $\left(\frac{n}{2}\right)$th and the $\left(\frac{n}{2}+1\right)$th observation if n is even

Example: Calculating the median

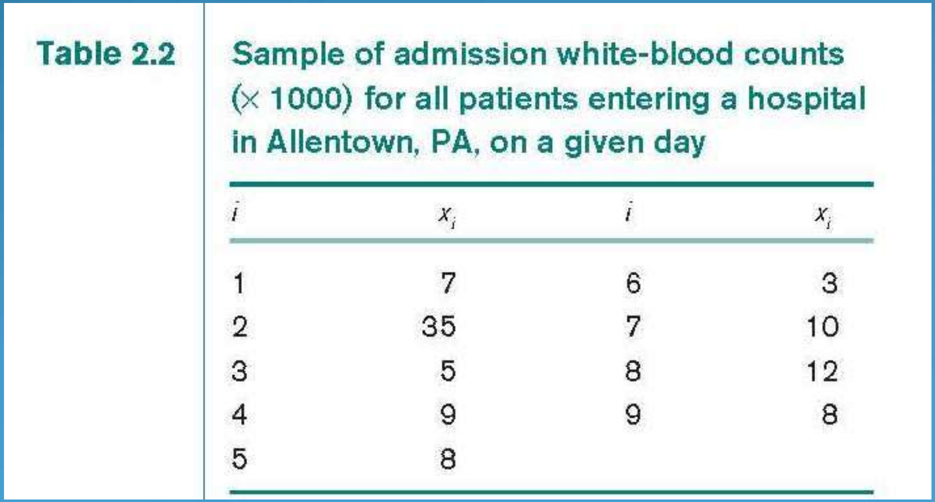


**Table 2.2** Sample of admission white-blood counts (× 1000) for all patients entering a hospital in Allentown, PA, on a given day

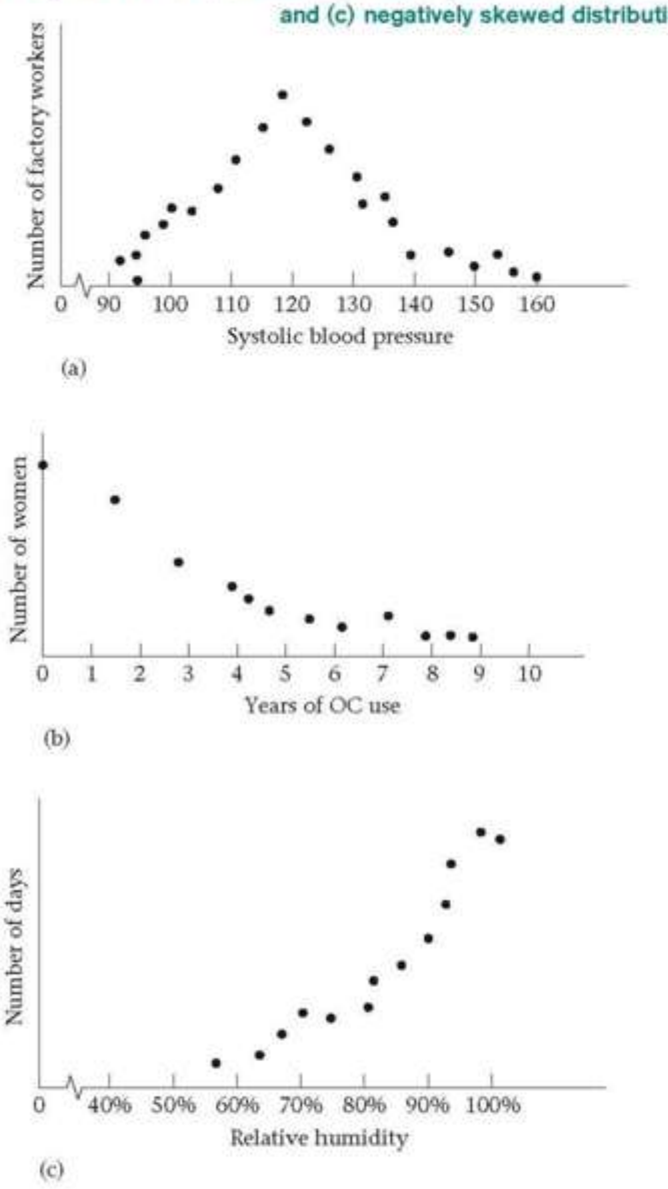
| $i$ | $x_i$ | $i$ | $x_i$ |
|---|---|---|---|
| 1 | 7 | 6 | 3 |
| 2 | 35 | 7 | 10 |
| 3 | 5 | 8 | 12 |
| 4 | 9 | 9 | 8 |
| 5 | 8 | | |

Order the sample as follows:
3, 5, 7, 8, 8, 9, 10, 12, 35

Because n is odd, the sample median is the fifth largest point, that is, 8

# Comparing Mean and Median

**Figure 2.3** **Graphic displays of (a) symmetric, (b) positively skewed, and (c) negatively skewed distributions**



For symmetric distributions, arithmetic mean is approximately the same as the median

For positively skewed distributions, arithmetic mean tends to be larger than the median

For negatively skewed distributions, the arithmetic mean tends to be smaller than the median

# Mode

Mode: the most frequently occurring value among all the observations in a sample.

Data distributions may have one or more modes.

One mode = unimodal

Two modes = bimodal

Three modes = trimodal and so on.

## Example

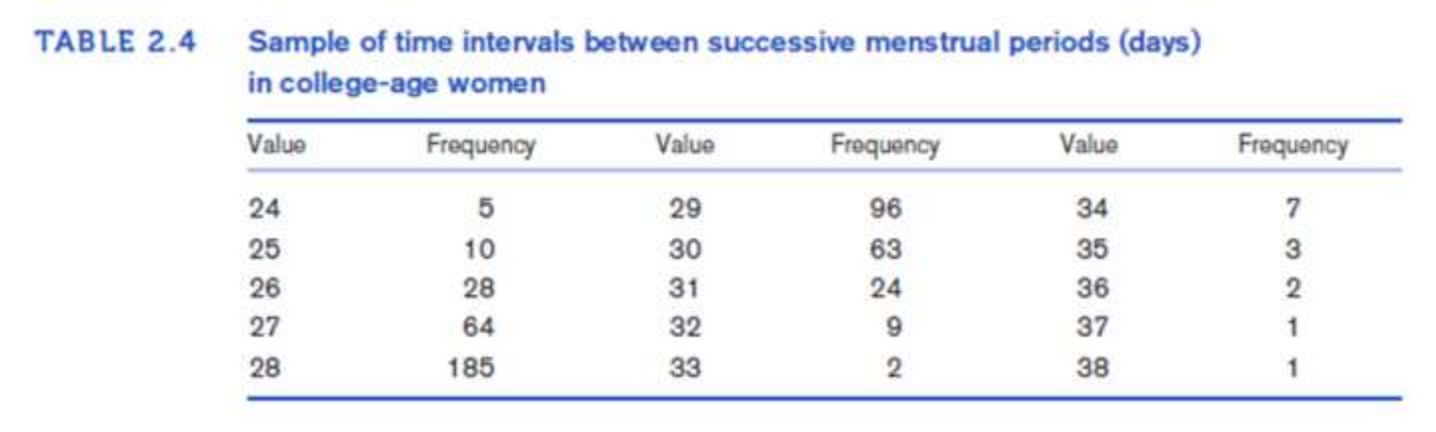


TABLE 2.4 Sample of time intervals between successive menstrual periods (days) in college-age women

| 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 |

Mode is 28

# Some Properties of Arithmetic Mean

Original sample : $x_1, \ldots, x_n$

Translated sample : $x_1 + c, \ldots, x_n + c$ (where c is some constant)

Let $y_i = x_i + c$ $\quad i = 1, \ldots, n$ then $\overline{y} = \overline{x} + c$

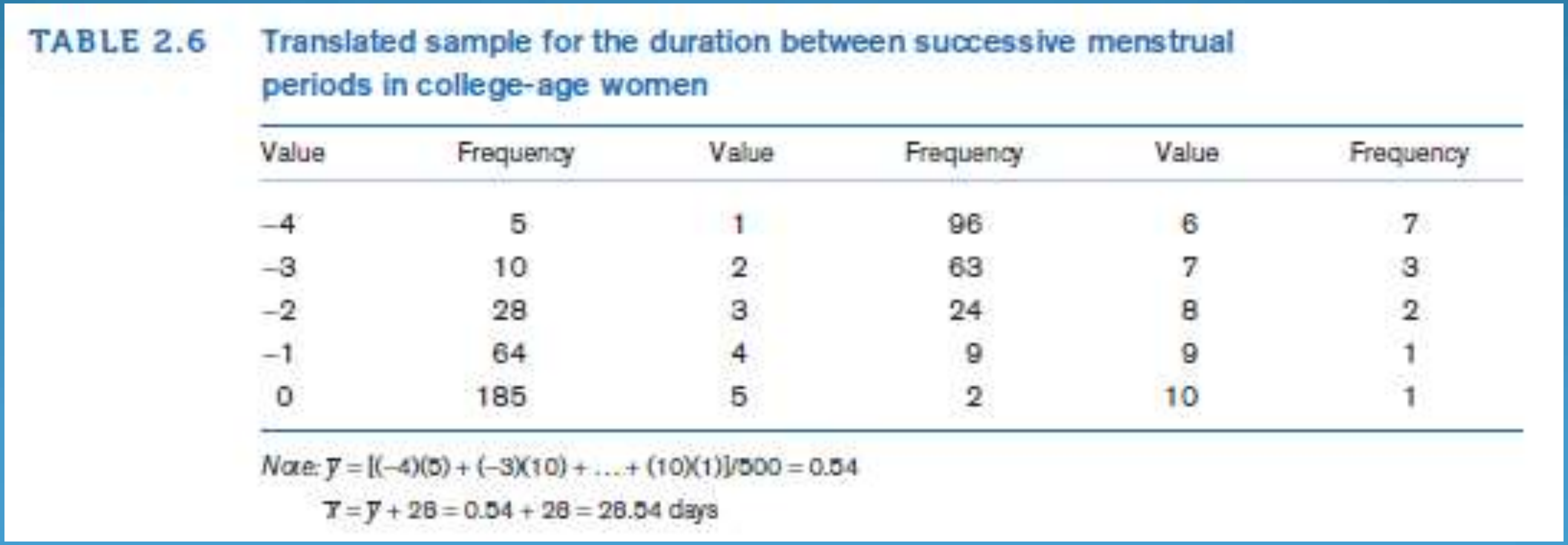


**TABLE 2.6** Translated sample for the duration between successive menstrual periods in college-age women

| Value | Frequency | Value | Frequency | Value | Frequency |
|---|---|---|---|---|---|
| –4 | 5 | 1 | 96 | 6 | 7 |
| –3 | 10 | 2 | 63 | 7 | 3 |
| –2 | 28 | 3 | 24 | 8 | 2 |
| –1 | 64 | 4 | 9 | 9 | 1 |
| 0 | 185 | 5 | 2 | 10 | 1 |

*Note:* $\overline{y} = [(-4)(5) + (-3)(10) + \ldots + (10)(1)]/500 = 0.54$

$\overline{x} = \overline{y} + 28 = 0.54 + 28 = 28.54$ days

If the unit or scale changes, then using the **rescaled sample**

$$y_i = cx_i \qquad i = 1, \ldots, n$$

Arithmetic mean is then $\overline{y} = c\overline{x}$

Let $x_1, \ldots, x_n$ be the original sample of data.

Let $y_i = c_1 x_i + c_2$ $\qquad i = 1, \ldots, n$ represent a transformed

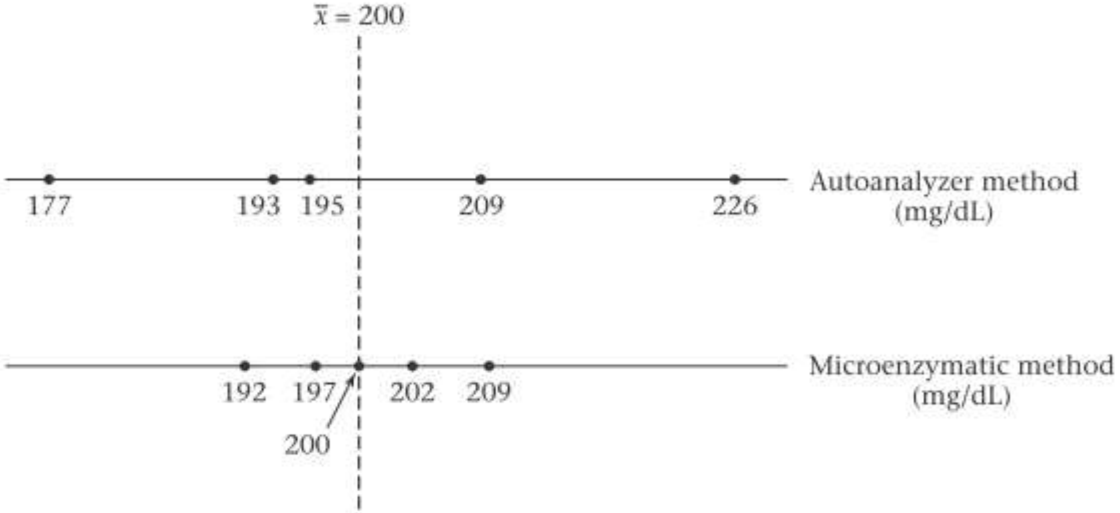
sample obtained by multiplying each original sample point by a factor $c_1$ and then shifting over by a constant $c_2$

If $y_i = c_1 x_i + c_2$ $\qquad i = 1, \ldots, n$

then $\overline{y} = c_1\overline{x} + c_2$

# Measures of Speed

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



The mean obtained by the two methods is the same. However, the **variability** or **spread** of the Autoanalyzer method appears to be greater.

# Range or variability

- Range is the difference between the largest and smallest observations in a sample.
- Once the sample is ordered, it is very easy to compute the range.
- Range is very sensitive to extreme observations or outliers.
- Larger the sample size (n), the larger the range and the more difficult the comparison between ranges from data sets of varying sizes.

A better approach to quantifying the spread in data sets is percentiles or quantiles.

Percentiles are less sensitive to outliers and are not greatly affected by the sample size.

# Quantiles or percentiles

The *p*th percentile is the value $V_p$ such that *p* percent of the sample points are less than or equal to $V_p$.

Median is the 50th percentile, which is a special case of a quantile.

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

- The (*k*+1)th largest sample point if *np*/100 is not an integer (where *k* is the largest integer less than *np*/100)
- The average of the (*np*/100)th and (*np*/100 +1)th largest observations if *np*/100 is an integer.

Frequently used percentiles are

- quartiles (25th, 50th, and 75th percentiles)
- quintiles (20th, 40th, 60th, and 80th percentiles)
- deciles (10th, 20th,…, 90th percentiles)

To compute percentiles, the sample points must be ordered.

If *n* is large, a stem-and-leaf plot or a computer program may be used.

# Variance and Standard Deviation

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 can be expressed as $x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x}$

that is, $d = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})}{n}$

The sum of the deviations of the individual observations of a sample about the sample mean is always zero.

The variance, which is the average of the squares of the deviations from the mean, given by is used as a measure of variation.

For a finite population with size N, the variance is defined as

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{N}$$

And for a sample of size n, the sample variance has the form

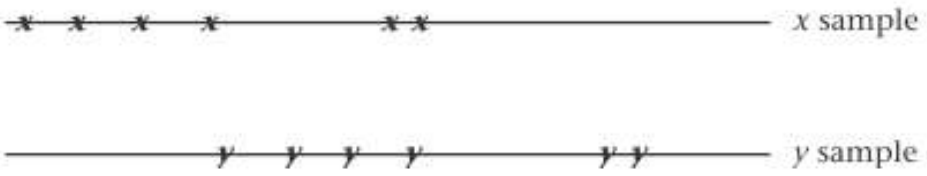
$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

To adjust the variance to the same measurement unit, the standard deviation, which is the positive square root of the variance, is commonly used

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

# Properties of Variance and Standard Deviation

**Figure 2.5** Comparison of the variances of two samples, where one sample has an origin shifted relative to the other



Samples $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ where $y_i = x_i + c$ $\quad i = 1, \ldots, n$

if respective sample variances are $s_x^2$ and $s_y^2$ then $s_y^2 = s_x^2$

Samples $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ where $y_i = cx_i$ $\quad i = 1, \ldots, n$ *and* $c>0$

then $s_y^2 = c^2 s_x^2$ which is $s_y = cs_x$

$$s_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = \frac{\sum_{i=1}^{n}(cx_i - c\bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n}[c(x_i - \bar{x})]^2}{n-1} = \frac{\sum_{i=1}^{n}c^2(x_i - \bar{x})^2}{n-1}$$

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

$$s_y = \sqrt{c^2 s_x^2} = cs_x$$

# Coefficient of Variation (CV)

Defined as 100% × $(s/\bar{x})$

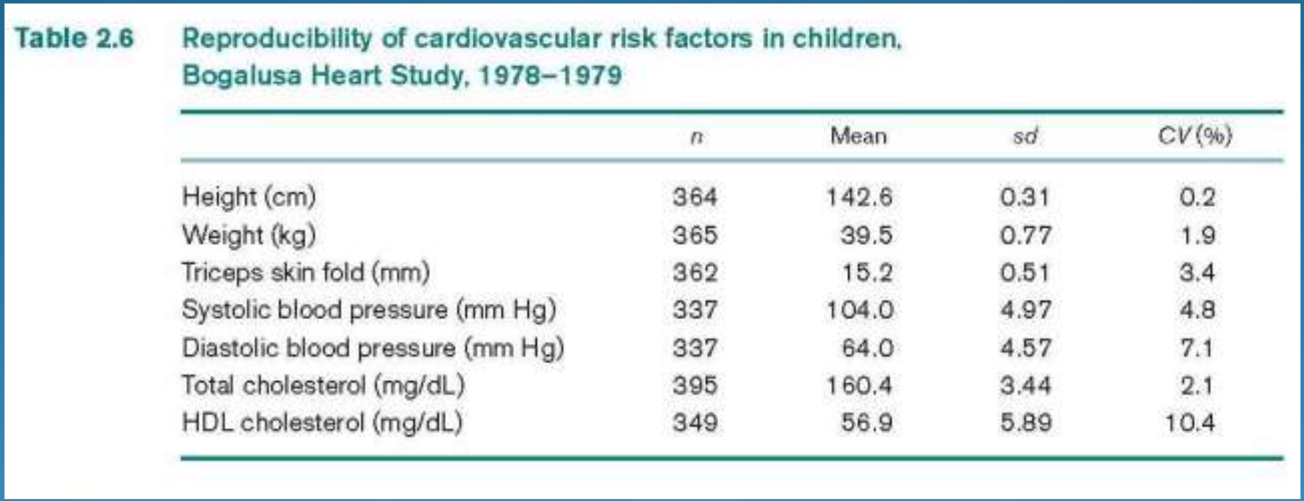


**Table 2.6** **Reproducibility of cardiovascular risk factors in children, Bogalusa Heart Study, 1978–1979**

| | *n* | Mean | *sd* | *CV* (%) |
|---|---|---|---|---|
| Height (cm) | 364 | 142.6 | 0.31 | 0.2 |
| Weight (kg) | 365 | 39.5 | 0.77 | 1.9 |
| Triceps skin fold (mm) | 362 | 15.2 | 0.51 | 3.4 |
| Systolic blood pressure (mm Hg) | 337 | 104.0 | 4.97 | 4.8 |
| Diastolic blood pressure (mm Hg) | 337 | 64.0 | 4.57 | 7.1 |
| Total cholesterol (mg/dL) | 395 | 160.4 | 3.44 | 2.1 |
| HDL cholesterol (mg/dL) | 349 | 56.9 | 5.89 | 10.4 |

Remains the same regardless of units used

Useful in comparing variability of different samples with different arithmetic means

Useful for comparing the reproducibility of different variables

# Grouped Data

When sample size is too large to display all the raw data, data are frequently collected in grouped form.

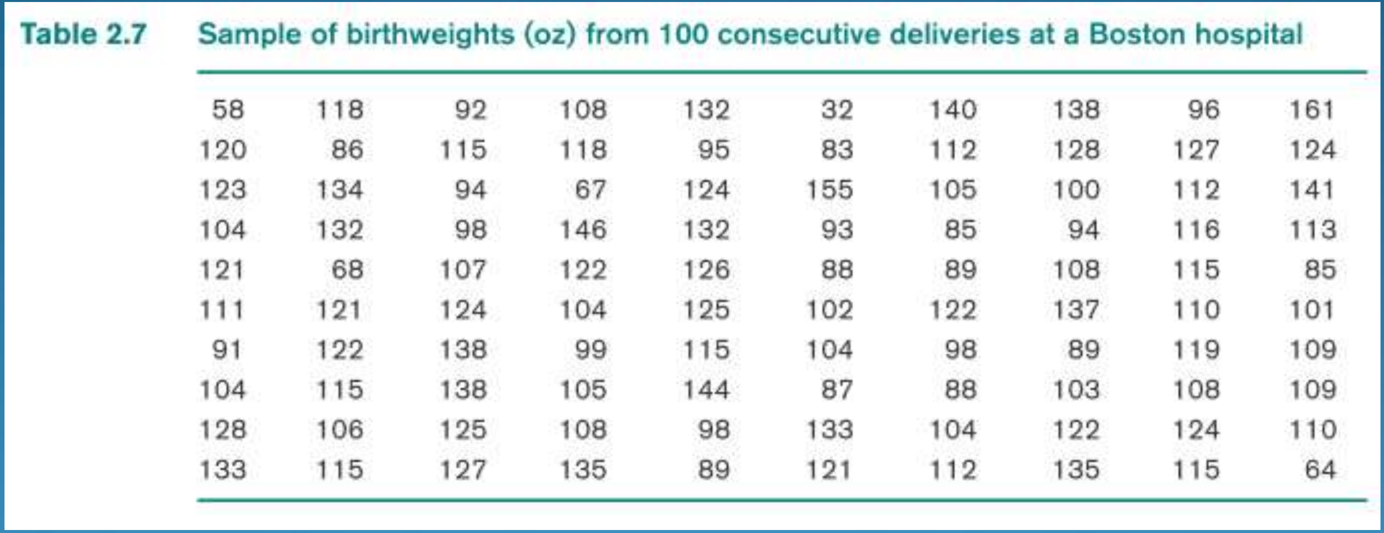
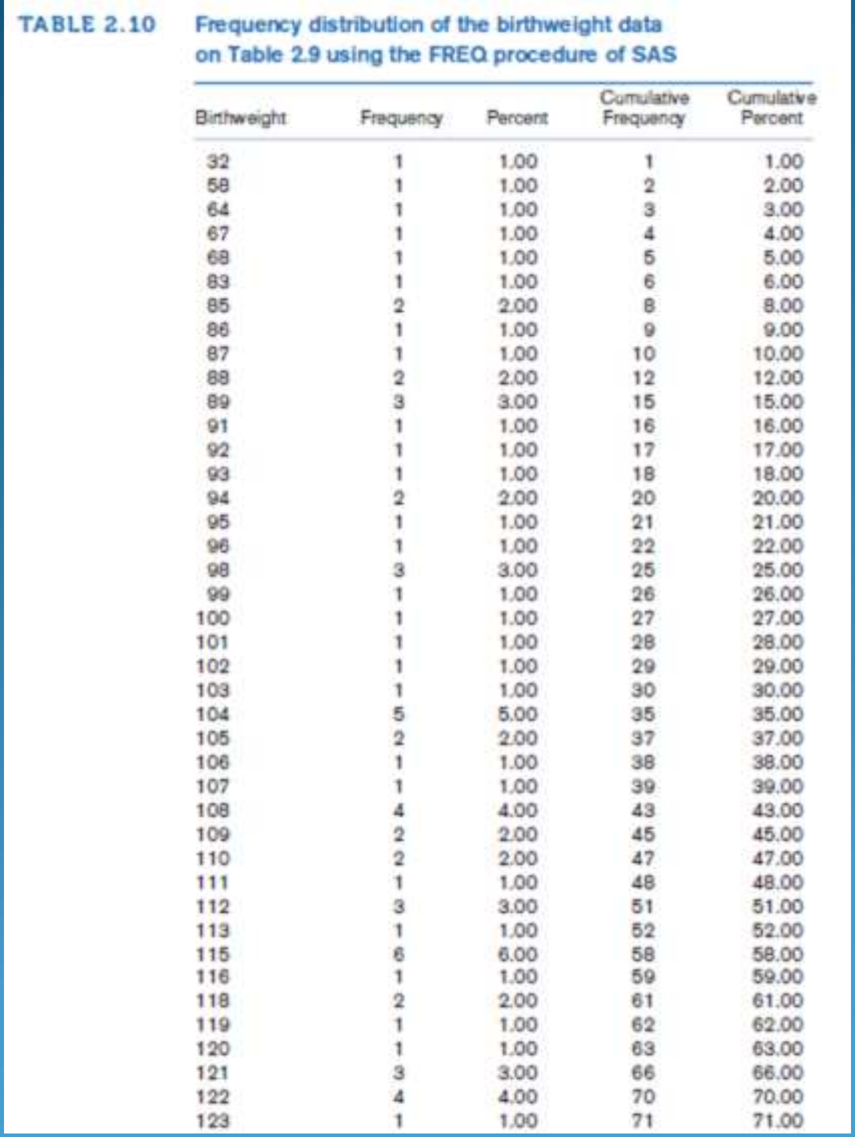


Table 2.7 Sample of birthweights (oz) from 100 consecutive deliveries at a Boston hospital

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 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 |

The simplest way to display the data is to generate a frequency distribution using a statistical package.

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.



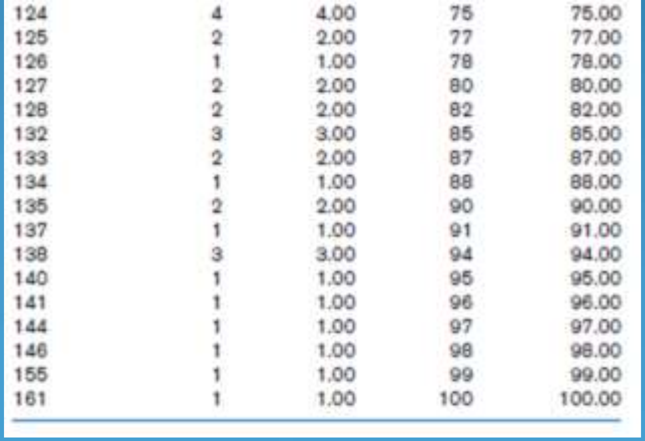


**TABLE 2.10** Frequency distribution of the birthweight data on Table 2.9 using the FREQ procedure of SAS

| 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 |

If the number of unique sample values is large, then a frequency distribution may still be too detailed.

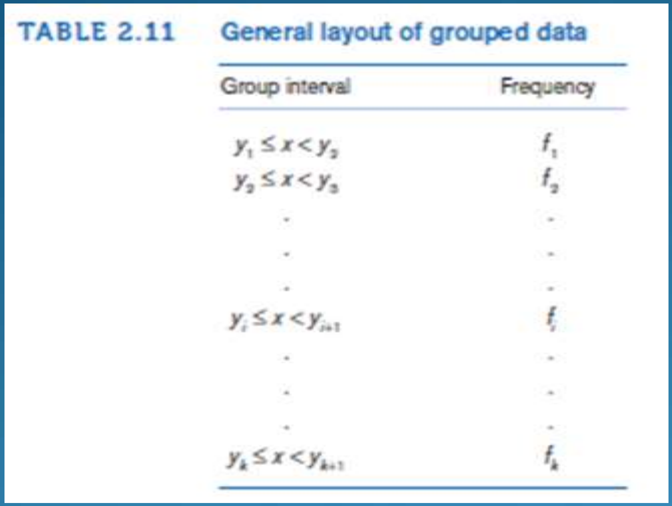


TABLE 2.11 General layout of grouped data

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

If the data is too large, then the data is categorized into broader groups.

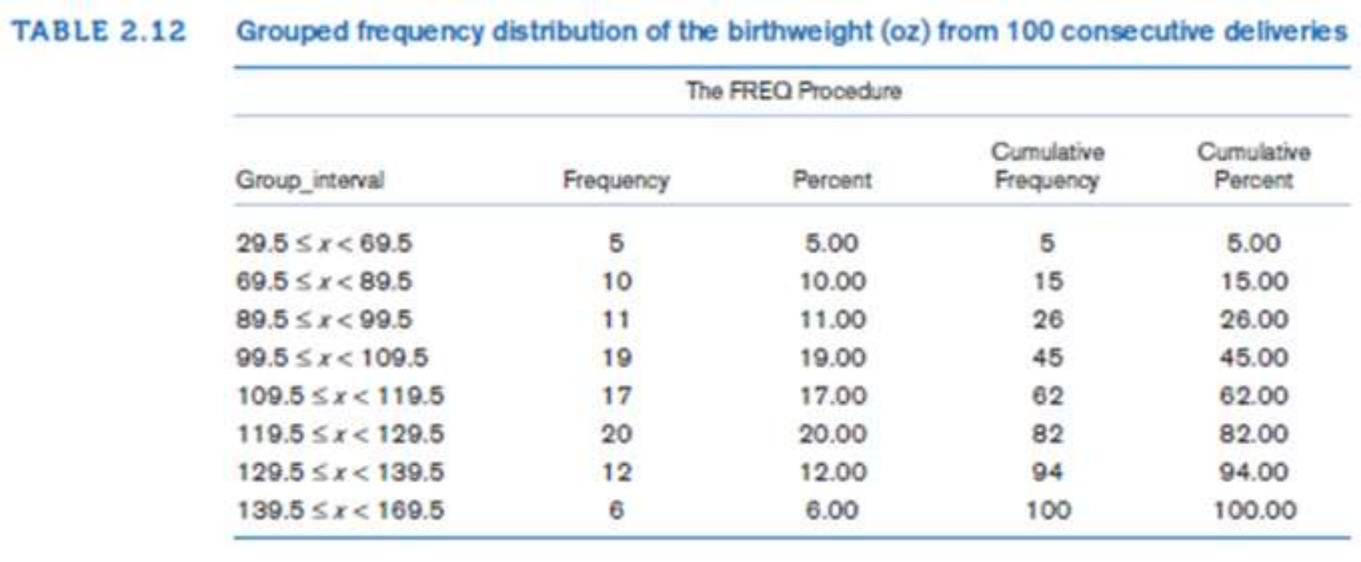


TABLE 2.12 Grouped frequency distribution of the birthweight (oz) from 100 consecutive deliveries

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 |

# Graphic Methods

Graphic methods of displaying data give a quick overall impression of data. The following are some graphic methods.

Bar graphs:

- ➢used to display grouped data;
- ➢difficult to construct;
- ➢Identity of the sample points within the respective groups is lost

Box plots:

- ➢Uses the relationships among the median, upper quantile, and lower quantile to describe the skewness or symmetry of a distribution

# Box plot

- If the distribution is symmetric, then upper and lower quartiles should be approximately equally spaced from the median

- If the upper quartile is farther from the median than the lower quartile, then the distribution is positively skewed

- If the lower quartile is farther from the median than the upper quartile, then the distribution is negatively skewed

# Outliers or Extreme Values

An outlying value is a value $x$ such that either

$x > \text{upper quartile} + 1.5 \times (\text{upper quartile} - \text{lower quartile})$

$x < \text{lower quartile} - 1.5 \times (\text{upper quartile} - \text{lower quartile})$

An extreme outlying value is a value $x$ such that either

$x > \text{upper quartile} + 3.0 \times (\text{upper quartile} - \text{lower quartile})$

$x < \text{lower quartile} - 3.0 \times (\text{upper quartile} - \text{lower quartile})$

- A vertical bar connects the upper quartile to the largest nonoutlying value in the sample
- A vertical bar connects the lower quartile to the smallest nonoutlying value in the sample

# Case study 1: Effects of lead exposure on neurological and psychological function in children

**Figure 2.9** Number of finger–wrist taps in the dominant hand for exposed and control groups, El Paso Lead Study

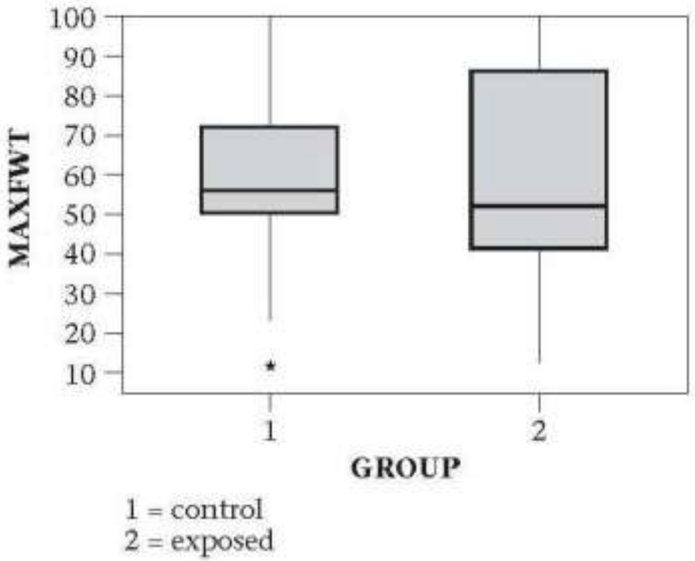
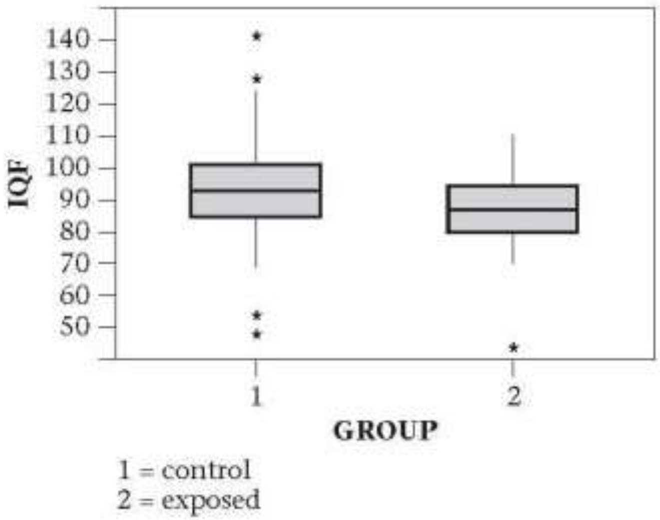


**Figure 2.10** Wechsler full-scale IQ scores for exposed and control groups, El Paso Lead Study



The finger-wrist tapping scores (MAXFWT) and full-scale IQ scores (IQF) seem slightly lower in the exposed group than in the control group.

# Obtaining descriptive statistics using a computer

- Numerous statistical packages may be used.

- Excel may be used to compute average (for the arithmetic mean), median (for the median), Stdev (for the standard deviation), Var (for the variance), and Percentile (for obtaining arbitrary percentiles from a sample). MINITAB has more features and statistical procedures.

# Summary

Numeric or graphic methods for displaying data help in

- quickly summarizing a data set
- And/or presenting results to others

A data set can be described numerically in terms of measure of location and a measure of spread

| Measure of location | Measure of spread |
|---|---|
| Arithmetic mean | Standard deviation |
| Median | Quantiles |
| Mode | Range |

Graphic methods include bar graphs and more exploratory methods such as box plots.

# The End