Chapter 4: Variability

Overview

- In statistics, our goal is to measure the amount of variability for a particular set of scores, a distribution.
- In simple terms, if the scores in a distribution are all the same, then there is no variability.
- If there are small differences between scores, then the variability is small, and if there are large differences between scores, then the variability is large.
- Definition: Variability provides a quantitative measure of the degree to which scores in a distribution are spread out or clustered together.

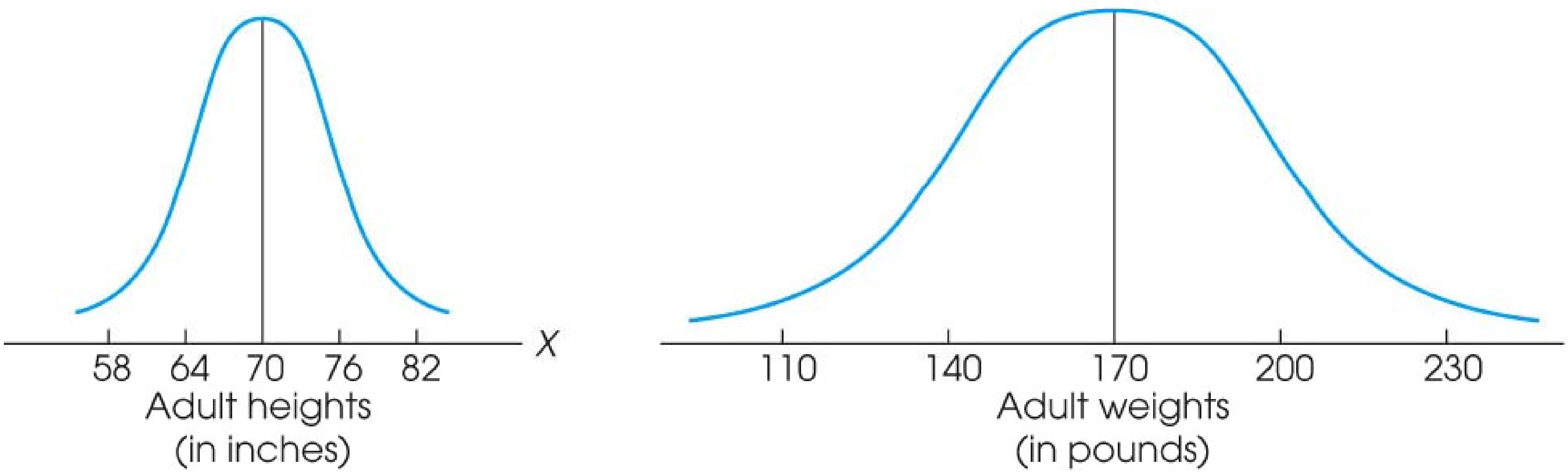

X
58
64
70
76
82
Adult heights
(in inches)
X
110
140
170
200
230
Adult weights
(in pounds)

Fig. 4-1, p. 106

Overview cont.

- In general, a good measure of variability serves two purposes:
  - Variability describes the distribution.
    - Specifically, it tells whether the scores are clustered close together or are spread out over a large distance.
  - Variability measures how well an individual score (or group of scores) represents the entire distribution.
    - This aspect of variability is very important for inferential statistics where relatively small samples are used to answer questions about populations.

Overview cont.

- In this chapter, we consider three different measures of variability:
  - Range
  - Interquartile Range
  - Standard Deviation.
- <u>Of these three, the standard deviation (and the related measure of variance) is by far the most important.</u>

The Range and Interquartile Range

- The range is the distance from the largest score to the smallest score in a distribution.
- Typically, the range is defined as the difference between the upper real limit of the largest X value and the lower real limit of the smallest X value.

For example, consider the following data:

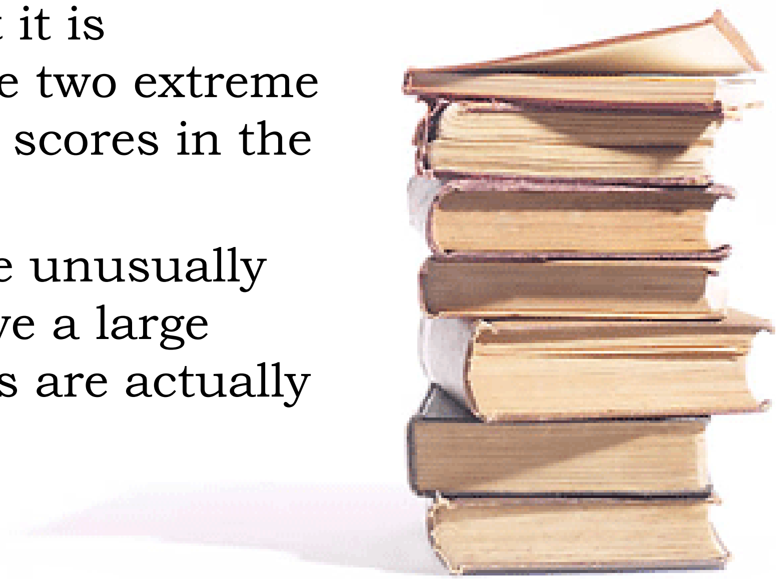
3, 7, 12, 8, 5, 10

For these data, $X_{max} = 12$, with an upper real limit of 12.5, and $X_{min} = 3$, with a lower real limit of 2.5. Thus, the range equals

$$\begin{aligned} \text{range} &= \text{URL } X_{max} - \text{LRL } X_{min} \\ &= 12.5 - 2.5 = 10 \end{aligned}$$

The Range cont.

- The range is perhaps the most obvious way of describing how spread out the scores are- simply find the distance between the maximum and the minimum scores.
- The problem with using the range as a measure of variability is that it is completely determined by the two extreme values and ignores the other scores in the distribution.
- Thus, a distribution with one unusually large (or small) score will have a large range even if the other scores are actually clustered close together.



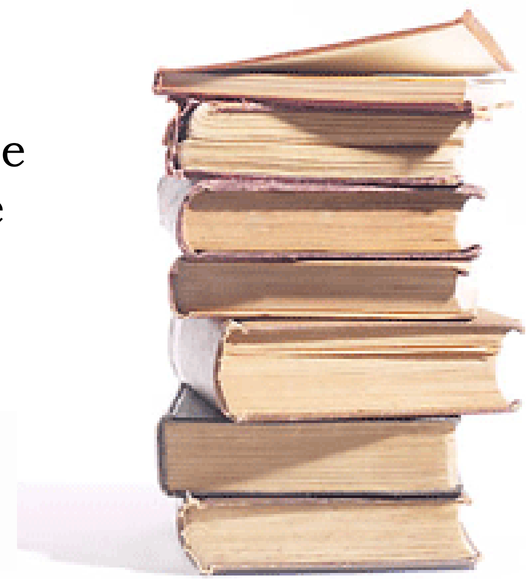
The Range cont.

- Because the range does not consider all the scores in the distribution, it often does not give an accurate description of the variability for the entire distribution.
- For this reason, the range is considered to be a crude and unreliable measure of variability.

The Interquartile Range

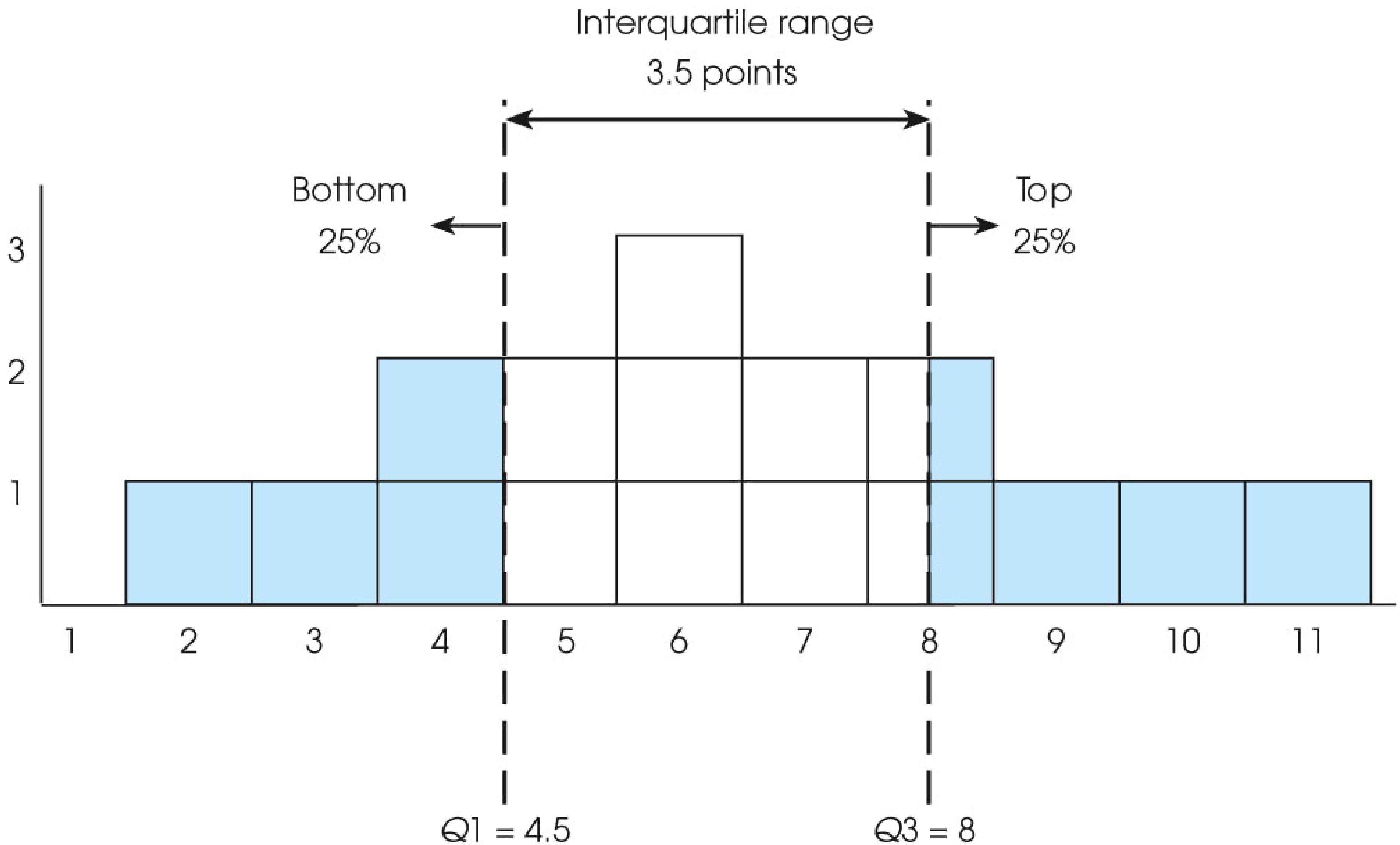
- One way to avoid the excessive influence of one or two extreme scores is to measure variability with the interquartile range.
- The interquartile range ignores extreme scores, instead, it measures the range covered by the middle 50% of the distribution.
- Definition: The interquartile range is the range covered by the middle 50% of the distribution.
  - Thus, the definitional formula is:

$$\text{interquartile range} = Q3 - Q1$$



The Interquartile Range cont.

- The simplest method for finding the values of Q1 and Q3 is to construct a frequency distribution histogram in which each score is represented by a box (Figure 4.2).
- When the interquartile range is used to describe variability, it commonly is transformed into the semi-interquartile range.
- As the name implies, the semi-interquartile range is one-half of the interquartile range.
- Conceptually, the semi-interquartile range measures the distance from the middle of the distribution to the boundaries that define the middle 50%.


Interquartile range
3.5 points
Bottom
25%
Top
25%
3
2
1
1
2
3
4
5
6
7
8
9
10
11
Q1 = 4.5
Q3 = 8

Semi-Interquartile Range

- The semi-interquartile range is half of the interquartile range:

$$\text{semi-interquartile range} = \frac{Q3 - Q1}{2}$$

- For the distribution in Figure 4.2 the interquartile range is 3.5 points. The semi-interquartile range is half of this distance:

$$\text{semi-interquartile range} = \frac{3.5}{2} = 1.75$$

- Because the semi-interquartile range (or interquartile range) is derived from the ntiddle 50% of a distribution, it is less likely to be influenced by extreme scores and therefore gives a better and more stable measure of variability than the range.

Semi-Interquartile Range cont.

- However, the semi-interquartile range only considers the middle 50% of the scores and completely disregards the other 50%.
- Therefore, it does not give a complete picture of the variability for the entire set of scores.
- <u>Like the range, the semi-interquartile range is considered to be a crude measure of variability</u>.

Standard Deviation and Variance for a Population

- <u>The standard deviation is the most commonly used and the most important measure of variability.</u>
- <u>Standard deviation uses the mean of the distribution as a reference point and measures variability by considering the distance between each score and the mean.</u>
- It determines whether the scores are generally near or far from the mean.
  - That is, are the scores clustered together or scattered?
  - In simple terms, the standard deviation approximates the average distance from the mean.

Standard Deviation and Variance for a Population cont.

- <u>Calculating the values</u>:
  - **<u>STEP 1</u>**: The first step in finding the standard distance from the mean is to determine the deviation, or distance from the mean, for each individual score. By definition, the deviation for each score is the difference between the score and the mean.
    - Definition: Deviation is distance from the mean:

$$\text{deviation score} = X - \mu$$

Standard Deviation and Variance for a Population cont.

- **STEP 2:** Because our goal is to compute a measure of the standard distance from the mean, the obvious next step is to calculate the mean of the deviation scores.
- To compute this mean, you first add up the deviation scores and then divide by N.
- This process is demonstrated in the following example.

We start with the following set of $N = 4$ scores. These scores add up to $\Sigma X = 12$, so the mean is $\mu = \frac{12}{4} = 3$. For each score, we have computed the deviation.

X	$X - \mu$
8	+5
1	−2
3	0
0	−3
	$0 = \Sigma(X - \mu)$

Standard Deviation and Variance for a Population cont.

- **<u>STEP 3:</u>** The average of the deviation scores will not work as a measure of variability because it is always zero.
- Clearly, this problem results from the positive and negative values canceling each other out.
- The solution is to get rid of the signs (+ and -).
- The standard procedure for accomplishing this is to square each deviation score.
- Using the squared values, you then compute the mean squared deviation, which is called variance.

Standard Deviation and Variance for a Population cont.

- Definition: Population variance equals the mean squared deviation. Variance is the average squared distance from the mean.
- **STEP 4:** Remember that our goal is to compute a measure of the standard distance from the mean.
- Variance, which measures the average squared distance from the mean, is not exactly what we want.
- The final step simply makes a correction for having squared all the distances.
  - The new measure, the *standard deviation*, is the square root of the *variance*.

$$\text{Standard deviation} = \sqrt{\text{variance}}$$

Standard Deviation and Variance for a Population cont.

- Because the standard deviation and variance are defined in terms of distance from the mean, these measures of variability are used only with numerical scores that are obtained from measurements on an interval or a ratio scale.

Formulas for Population Variance and Standard Deviation

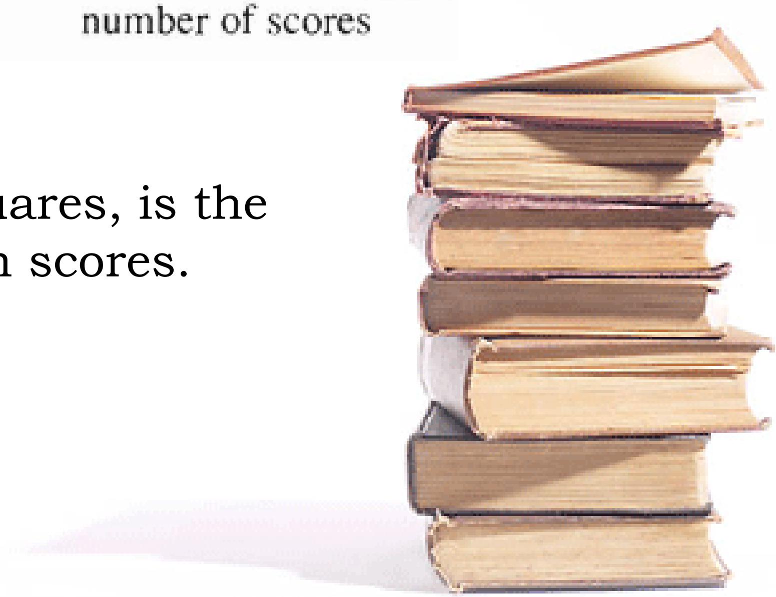
- The concepts of standard deviation and variance are the same for both samples and populations.
- However, the details of the calculations differ slightly, depending on whether you have data from a sample or from a complete population.
- We first consider the formulas for populations and then look at samples in Section 4.4.
- The sum of squared deviations (SS) Recall that variance is defined as the mean of the squared deviations.

Formulas for Population Variance and Standard Deviation cont.

- This mean is computed exactly the same way you compute any mean: First find the sum, and then divide by the number of scores.

$$\text{Variance} = \text{mean squared deviation} = \frac{\text{sum of squared deviations}}{\text{number of scores}}$$

- <u>Definition</u>: SS, or sum of squares, is the sum of the squared deviation scores.



Formulas for Population Variance and Standard Deviation cont.

- You will need to know two formulas to compute SS.
- These formulas are algebraically equivalent (they always produce the same answer), but they look different and are used in different situations.
- The first of these formulas is called the definitional formula because the terms in the formula literally define the process of adding up the squared deviations:

definitional formula: $SS = \Sigma(X - \mu)^2$

Formulas for Population Variance and Standard Deviation cont.

- Following the proper order of operations (page 25), the formula instructs you to perform the following sequence of calculations:

1. Find each deviation score $(X - \mu)$.
2. Square each deviation score, $(X - \mu)^2$.
3. Add the squared deviations.

Final Formulas and Notation

- With the definition and calculation of SS behind you, the equations for variance and standard deviation become relatively simple.
- Remember that variance is defined as the mean squared deviation.
- The mean is the sum divided by *N*, so the equation for the population variance is:

$$\text{variance} = \frac{SS}{N}$$

Final Formulas and Notation cont.

- Standard deviation is the square root of variance, so the equation for the population standard deviation is:

$$\text{standard deviation} = \sqrt{\frac{SS}{N}}$$

Final Formulas and Notation cont.

- Using the definitional formula for SS, the complete calculation of population variance can be expressed as:

$$\text{population variance} = \sigma^2 = \frac{\Sigma(X - \mu)^2}{N}$$

- However, the population variance is expressed as:

$$\text{population variance} = \sigma^2 = \frac{SS}{N}$$

Graphic Representation of the Mean and Standard Deviation

- In frequency distribution graphs, we identify the position of the population mean by drawing a vertical line and labeling it with μ (Figure 4.5).
- Because the *standard deviation* measures distance from the *mean*, it will be represented by a line or an arrow drawn from the mean outward for a distance equal to the *standard deviation* (see Figure 4.5).
- <u>You should realize that we could have drawn the arrow pointing to the left, or we could have drawn two arrows, with one pointing to the right and one pointing to the left.</u>
- In each case, the goal is to show the standard distance from the *mean*.

FIGURE 4.5

The graphic representation of a population with a mean of $\mu = 40$ and a standard deviation of $\mu = 4$.

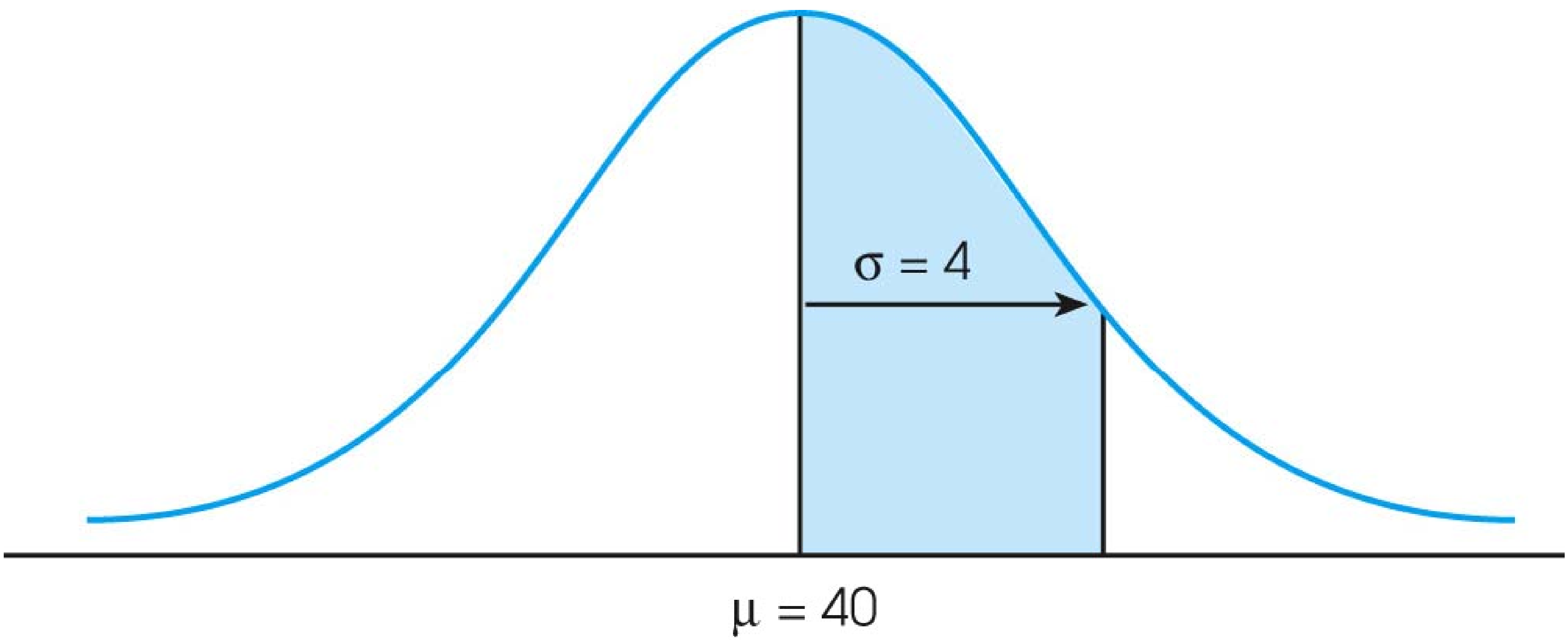


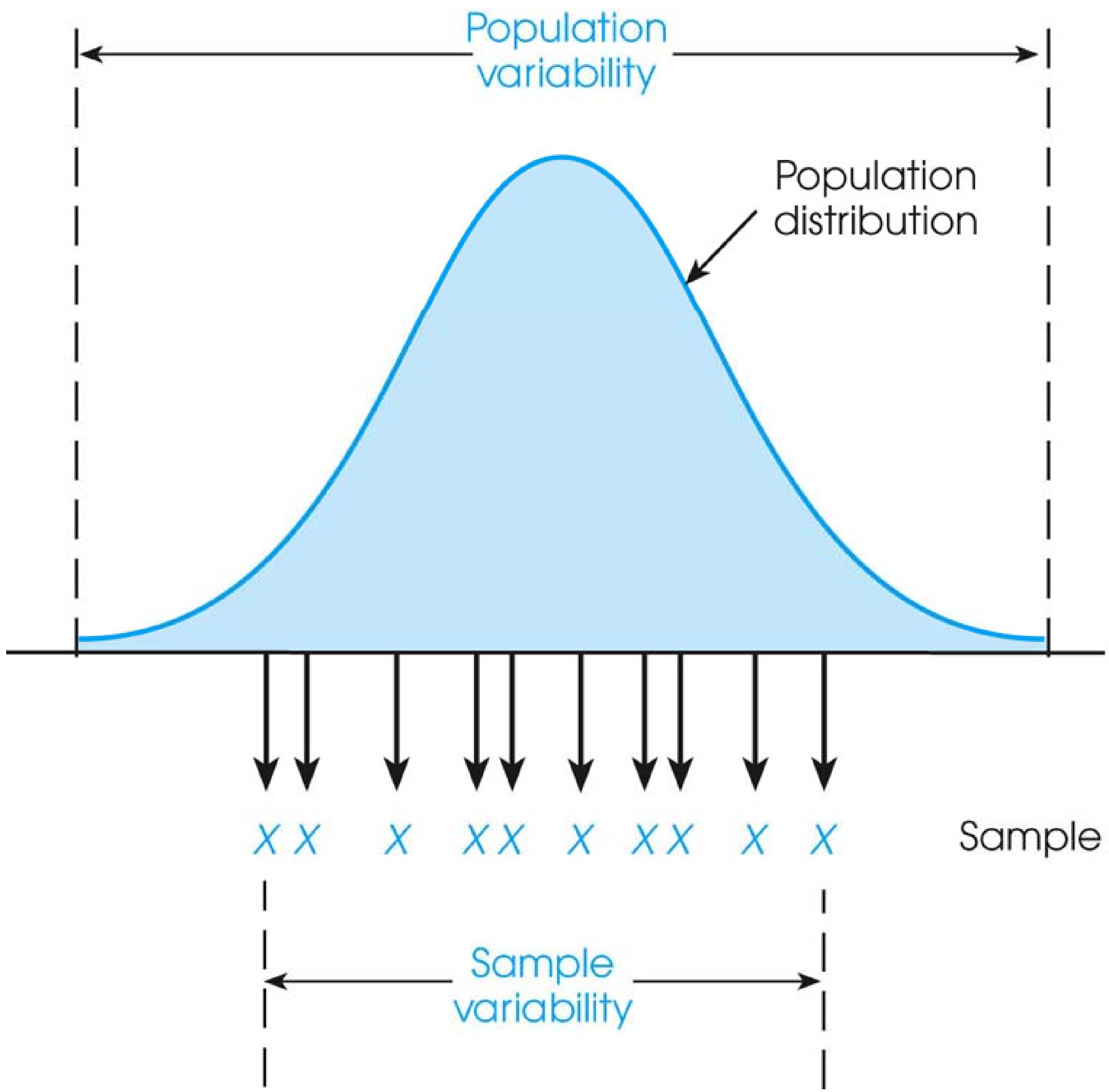
Fig. 4-5, p. 116

Standard Deviation and Variances for Samples

- The goal of inferential statistics is to use the limited information from samples to draw general conclusions about populations.
- The basic assumption of this process is that samples should be representative of the populations from which they come.
- This assumption poses a special problem for variability because samples consistently tend to be less variable than their populations.
- The fact that a sample tends to be less variable than its population means that sample variability gives a biased estimate of population variability.

Standard Deviation and Variances for Samples cont.

- This bias is in the direction of underestimating the population value rather than being right on the mark.
- Please see example on next slide.

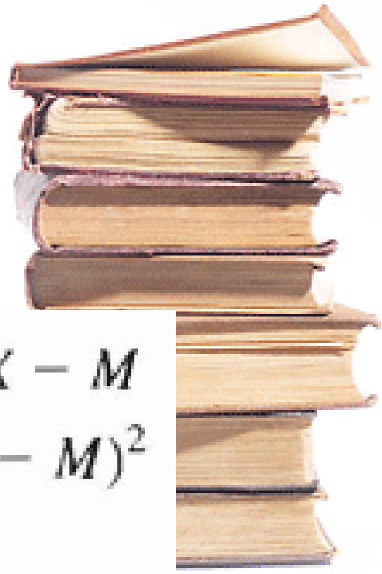


Population
variability
Population
distribution
X X X X X X X X X X
Sample
Sample
variability

Fig. 4-6, p. 117

Standard Deviation and Variances for Samples cont.

- Fortunately, the bias in sample variability is consistent and predictable, which means it can be corrected.
- The calculations of *variance* and *standard deviation* for a sample follow the same steps that were used to find population variance and standard deviation.
- Except for minor changes in notation, the first three steps in this process are exactly the same for a sample as they were for a population.



1. Find the deviation for each score: deviation $= X - M$
2. Square each deviation: squared deviation $= (X - M)^2$
3. Add the squared deviations: $SS = \Sigma(X - M)^2$

Standard Deviation and Variances for Samples cont.

These three steps can be summarized in a definitional formula for *SS*:

$$\text{Definitional formula: } SS = \Sigma(X - M)^2$$

- Again, calculating *SS* for a sample is exactly the same as for a population, except for minor changes in notation.
- After you compute *SS*, however, it becomes critical to differentiate between samples and populations.
- To correct for the bias in sample variability, it is necessary to make an adjustment in the formulas for sample *variance* and *standard deviation*.

Standard Deviation and Variances for Samples cont.

- With this in mind, sample *variance* (identified by the symbol S^2) is defined as:

$$\text{sample variance} = s^2 = \frac{SS}{n-1}$$

- Using the definitional formula for *SS*. The complete calculation of sample *variance* can be expressed as:

$$\text{sample variance} = s^2 = \frac{\Sigma(X-M)^2}{n-1}$$

Standard Deviation and Variances for Samples cont.

- Sample *standard deviation* (identified by the symbol s) is simply the square root of the *variance*.

$$\text{sample standard deviation} = s = \sqrt{s^2} = \sqrt{\frac{SS}{n-1}}$$

- Notice that these sample formulas use (n-1) instead of n.
- This is the adjustment that is necessary to correct for the bias in sample variability.
- Remember that the formulas for sample *variance* and *standard deviation* were constructed so that the sample variability would provide a good estimate of population variability.

Standard Deviation and Variances for Samples cont.

- For this reason, the sample variance is often called *estimated population variance*, and the sample standard deviation is called *estimated population standard deviation.*
- <u>When you have only a sample to work with, the sample variance and standard deviation provide the best possible estimates of the population variability</u>.

Degrees of Freedom

- For a sample of n scores, the *degrees of freedom* or *df* for the sample variance are defined as *df* = n-1.
- The *degrees of freedom* determine the number of scores in the sample that are independent and free to vary.