Chapter 3

Data Description

Outline

- 3-1 Introduction
- 3-2 Measures of Central Tendency
- 3-3 Measures of Variation
- 3-4 Measures of Position
- 3-5 Exploratory Data Analysis

Objectives

- **Summarize data using the measures of central tendency, such as the mean, median, mode, and midrange.**
- **Describe data using the measures of variation, such as the range, variance, and standard deviation.**

Objectives

- Identify the position of a data value in a data set using various measures of position, such as percentiles, deciles and quartiles.

Objectives

- Use the techniques of exploratory data analysis, including stem and leaf plots, box plots, and five-number summaries to discover various aspects of data.

3-2 Measures of Central Tendency

- A **statistic** is a characteristic or measure obtained by using the data values from a sample.
- A **parameter** is a characteristic or measure obtained by using the data values from a specific population.

3-2 The Mean (arithmetic average)

- The **mean** is defined to be the sum of the data values divided by the total number of values.
- We will compute two means: one for the sample and one for a finite population of values.

3-2 The Mean (arithmetic average)

- **The mean, in most cases, is not an actual data value.**

3-2 The Sample Mean

The symbol $\overline{X}$ *represents the sample mean.*
$\overline{X}$ *is read as "*X*-bar". The Greek symbol*
$\sum$ *is read as "sigma" and it means "to sum".*

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

$$= \frac{\sum X}{n}.$$

3-2 The Sample Mean - Example

The ages in weeks of a random sample of six kittens at an animal shelter are 3, 8, 5, 12, 14, *and* 12. *Find the average age of this sample.*

The sample mean is

$$\overline{X} = \frac{\sum X}{n} = \frac{3+8+5+12+14+12}{6}$$

$$= \frac{54}{6} = 9 \textit{ weeks}.$$

3-2 The Population Mean

The Greek symbol μ represents the population mean. The symbol μ is read as "mu".
N is the size of the finite population.

$$\mu = \frac{X_1 + X_2 + \ldots + X_N}{N}$$
$$= \frac{\sum X}{N}.$$

3-2 The Population Mean - Example

A small company consists of the owner, the manager, the salesperson, and two technicians. The salaries are listed as \$50,000, 20,000, 12,000, 9,000 *and* 9,000 *respectively.* (*Assume this is the population.*)

Then the population mean will be

$$\mu = \frac{\sum X}{N}$$

$$= \frac{50{,}000 + 20{,}000 + 12{,}000 + 9{,}000 + 9{,}000}{5}$$

$$= \$20{,}000.$$

3-2 The Sample Mean for an Ungrouped Frequency Distribution

The mean for an ungrouped frequency distributuion is given by

$$\overline{X} = \frac{\sum(f \cdot X)}{n}.$$

Here f is the frequency for the corresponding value of X, *and* $n = \sum f$.

3-2 The Sample Mean for an Ungrouped Frequency Distribution - Example

The scores for 25 students on a 4 – point quiz are given in the table. Find the mean score.

Score, X	*Frequency, f*
0	*2*
1	*4*
2	*12*
3	*4*
4	*3*

3-2 The Sample Mean for an Ungrouped Frequency Distribution - Example

Score, X	Frequency, f	$f \cdot X$
0	2	0
1	4	4
2	12	24
3	4	12
4	3	12

$$\bar{X} = \frac{\sum f \cdot X}{n} = \frac{52}{25} = 2.08.$$

3-2 The Sample Mean for a Grouped Frequency Distribution

The mean for a grouped frequency distributuion is given by

$$\bar{X} = \frac{\sum (f \cdot X_m)}{n}.$$

Here X_m is the corresponding class midpoint.

3-2 The Sample Mean for a Grouped Frequency Distribution - Example

Given the table below, find the mean.

Class	Frequency, f
15.5 - 20.5	3
20.5 - 25.5	5
25.5 - 30.5	4
30.5 - 35.5	3
35.5 - 40.5	2

3-2 The Sample Mean for a Grouped Frequency Distribution - Example

Table with class midpoints, X_m.

Class	*Frequency, f*	X_m	$f \cdot X_m$
15.5 - 20.5	*3*	*18*	*54*
20.5 - 25.5	*5*	*23*	*115*
25.5 - 30.5	*4*	*28*	*112*
30.5 - 35.5	*3*	*33*	*99*
35.5 - 40.5	*2*	*38*	*76*

3-2 The Sample Mean for a Grouped Frequency Distribution - Example

$$\sum f \cdot X_m = 54 + 115 + 112 + 99 + 76$$

$$= 456$$

$$and\ n = 17. \quad So$$

$$\overline{X} = \frac{\sum f \cdot X_m}{n}$$

$$= \frac{456}{17} = 26.82.$$

3-2 The Median

- When a data set is ordered, it is called a data array.
- The median is defined to be the midpoint of the data array.
- The symbol used to denote the median is MD.

3-2 The Median - Example

- The weights (in pounds) of seven army recruits are 180, 201, 220, 191, 219, 209, and 186. Find the median.
- Arrange the data in order and select the middle point.

3-2 The Median - Example

- Data array: 180, 186, 191, 201, 209, 219, 220.
- The median, MD = 201.

3-2 The Median

- In the previous example, there was an odd number of values in the data set. In this case it is easy to select the middle number in the data array.

3-2 The Median

- When there is an even number of values in the data set, the median is obtained by taking the average of the two middle numbers.

3-2 The Median - Example

- Six customers purchased the following number of magazines: 1, 7, 3, 2, 3, 4. Find the median.
- Arrange the data in order and compute the middle point.
- Data array: 1, 2, 3, 3, 4, 7.
- The median, MD = (3 + 3)/2 = 3.

3-2 The Median - Example

- The ages of 10 college students are: 18, 24, 20, 35, 19, 23, 26, 23, 19, 20. Find the median.
- Arrange the data in order and compute the middle point.

3-2 The Median - Example

- Data array: 18, 19, 19, 20, 20, 23, 23, 24, 26, 35.
- The median, MD = (20 + 23)/2 = 21.5.

3-2 The Median-Ungrouped Frequency Distribution

- For an ungrouped frequency distribution, find the median by examining the cumulative frequencies to locate the middle value.

3-2 The Median-Ungrouped Frequency Distribution

- If n is the sample size, compute $n/2$. Locate the data point where $n/2$ values fall below and $n/2$ values fall above.

3-2 The Median-Ungrouped Frequency Distribution - Example

- LRJ Appliance recorded the number of VCRs sold per week over a one-year period. The data is given below.

No. Sets Sold	*Frequency*
1	*4*
2	*9*
3	*6*
4	*2*
5	*3*

3-2 The Median-Ungrouped Frequency Distribution - Example

- To locate the middle point, divide *n* by 2; 24/2 = 12.
- Locate the point where 12 values would fall below and 12 values will fall above.
- Consider the cumulative distribution.
- The 12th and 13th values fall in class 2. Hence MD = 2.

3-2 The Median-Ungrouped Frequency Distribution - Example

No. Sets Sold	Frequency	Cumulative Frequency
1	4	4
2	9	13
3	6	19
4	2	21
5	3	24

This class contains the 5th through the 13th values.

3-2 The Median for a Grouped Frequency Distribution

The median can be computed from:

$$MD = \frac{(n/2) - cf}{f}(w) + L_m$$

Where

n = *sum of the frequencies*

cf = *cumulative frequency of the class immediately preceding the median class*

f = *frequency of the median class*

w = *width of the median class*

L_m = *lower boundary of the median class*

3-2 The Median for a Grouped Frequency Distribution - Example

Given the table below, find the median.

Class	*Frequency, f*
15.5 - 20.5	3
20.5 - 25.5	5
25.5 - 30.5	4
30.5 - 35.5	3
35.5 - 40.5	2

3-2 The Median for a Grouped Frequency Distribution - Example

Table with cumulative frequencies.

Class	Frequency, f	Cumulative Frequency
15.5 - 20.5	3	3
20.5 - 25.5	5	8
25.5 - 30.5	4	12
30.5 - 35.5	3	15
35.5 - 40.5	2	17

3-2 The Median for a Grouped Frequency Distribution - Example

- To locate the halfway point, divide *n* by 2; $17/2 = 8.5 \approx 9$.
- Find the class that contains the 9th value. This will be the median class.
- Consider the cumulative distribution.
- The median class will then be 25.5 – 30.5.

3-2 The Median for a Grouped Frequency Distribution

$$n = 17$$

$$cf = 8$$

$$f = 4$$

$$w = 25.5 - 20.5 = 5$$

$$L_m = 25.5$$

$$MD = \frac{(n/2) - cf}{f}(w) + L_m = \frac{(17/2) - 8}{4}(5) + 25.5$$

$$= 26.125.$$

3-2 The Mode

- The mode is defined to be the value that occurs most often in a data set.
- A data set can have more than one mode.
- A data set is said to have no mode if all values occur with equal frequency.

3-2 The Mode - Examples

- The following data represent the duration (in days) of U.S. space shuttle voyages for the years 1992-94. Find the mode.
- Data set: 8, 9, 9, 14, 8, 8, 10, 7, 6, 9, 7, 8, 10, 14, 11, 8, 14, 11.
- Ordered set: 6, 7, 7, 8, 8, 8, 8, 8, 9, 9, 9, 10, 10, 11, 11, 14, 14, 14. Mode = 8.

3-2 The Mode - Examples

- **Six strains of bacteria were tested to see how long they could remain alive outside their normal environment. The time, in minutes, is given below. Find the mode.**
- **Data set: 2, 3, 5, 7, 8, 10.**
- **There is no mode since each data value occurs equally with a frequency of one.**

3-2 The Mode - Examples

- Eleven different automobiles were tested at a speed of 15 mph for stopping distances. The distance, in feet, is given below. Find the mode.
- Data set: 15, 18, 18, 18, 20, 22, 24, 24, 24, 26, 26.
- There are two modes (bimodal). The values are 18 and 24. *Why?*

3-2 The Mode for an Ungrouped Frequency Distribution - Example

Given the table below, find the mode.

Values	*Frequency, f*
15	*3*
20	*5*
25	*8*
30	*3*
35	*2*

Mode

3-2 The Mode - Grouped Frequency Distribution

- The mode for grouped data is the modal class.
- The modal class is the class with the largest frequency.
- Sometimes the midpoint of the class is used rather than the boundaries.

3-2 The Mode for a Grouped Frequency Distribution - Example

Given the table below, find the mode.

Modal Class

Class	*Frequency, f*
15.5 - 20.5	*3*
20.5 - 25.5	*5*
25.5 - 30.5	*7*
30.5 - 35.5	*3*
35.5 - 40.5	*2*

3-2 The Midrange

- The midrange is found by adding the lowest and highest values in the data set and dividing by 2.
- The midrange is a rough estimate of the middle value of the data.
- The symbol that is used to represent the midrange is MR.

3-2 The Midrange - Example

- Last winter, the city of Brownsville, Minnesota, reported the following number of water-line breaks per month. The data is as follows: 2, 3, 6, 8, 4, 1. Find the midrange. MR = (1 + 8)/2 = 4.5.
- *Note:* Extreme values influence the midrange and thus may not be a typical description of the middle.

3-2 The Weighted Mean

- The **weighted mean** is used when the values in a data set are not all equally represented.
- The **weighted mean of a variable *X*** is found by multiplying each value by its corresponding weight and dividing the sum of the products by the sum of the weights.

3

3-2 The Weighted Mean

The weighted mean

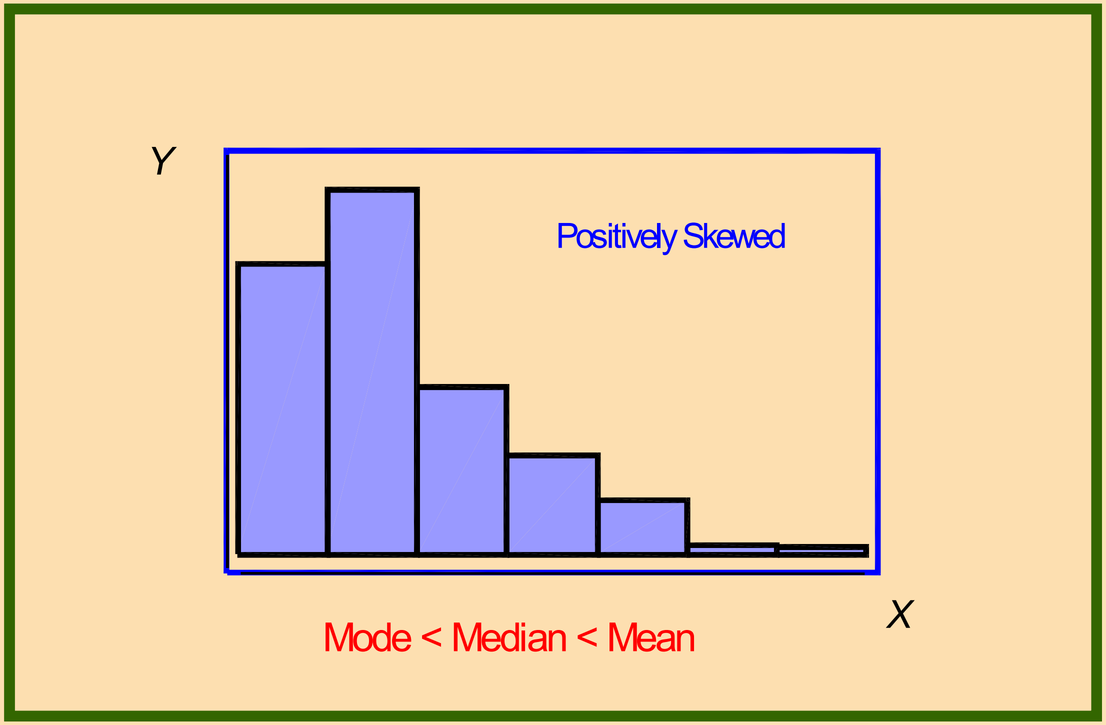
$$\overline{X} = \frac{w_1 X_1 + w_2 X_2 + \ldots + w_n X_n}{w_1 + w_2 + \ldots + w_n} = \frac{\sum wX}{\sum w}$$

where $w_1, w_2, \ldots, w_n$ *are the weights for the values* $X_1, X_2, \ldots, X_n$.

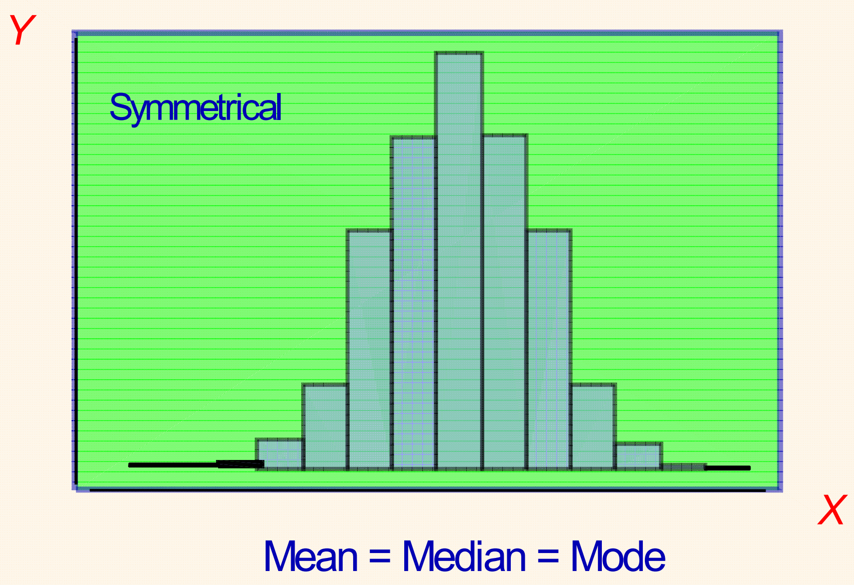
Distribution Shapes

- Frequency distributions can assume many shapes.
- The three most important shapes are positively skewed, symmetrical, and negatively skewed.

Positively Skewed

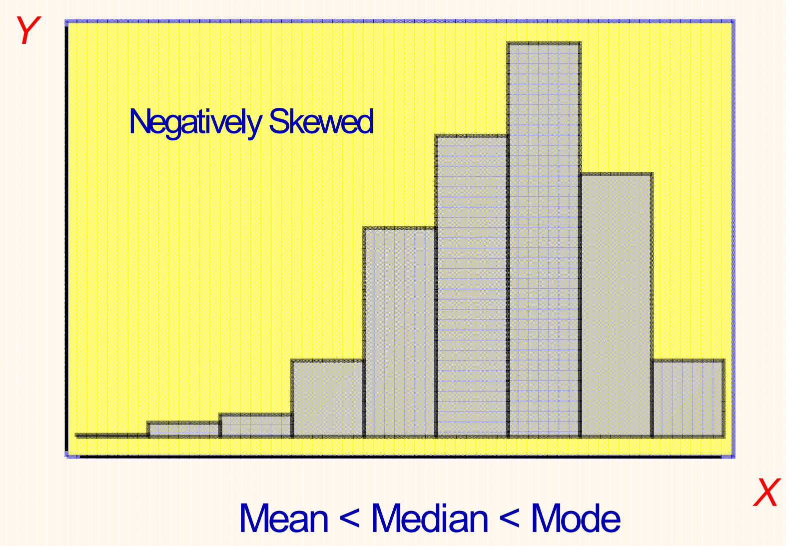


Symmetrical



2

Negatively Skewed



3-3 Measures of Variation - Range

- The range is defined to be the highest value minus the lowest value. The symbol R is used for the range.
- R = highest value – lowest value.
- Extremely large or extremely small data values can drastically affect the range.

3-54

3-3 Measures of Variation - Population Variance

The variance is the average of the squares of the distance each value is from the mean.

The symbol for the population variance is σ^2 (σ is the Greek lowercase letter sigma)

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}, \text{ where}$$

X = individual value

μ = population mean

N = population size

3-3 Measures of Variation - Population Standard Deviation

The standard deviation is the square root of the variance.

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum (X - \mu)^2}{N}}.$$

3-3 Measures of Variation - Example

- **Consider the following data to constitute the population: 10, 60, 50, 30, 40, 20. Find the mean and variance.**
- **The mean μ = (10 + 60 + 50 + 30 + 40 + 20)/6 = 210/6 = 35.**
- **The variance σ^2 = 1750/6 = 291.67.** **See next slide for computations.**

3-3 Measures of Variation - Example

X	$X - \mu$	$(X - \mu)^2$
10	-25	625
60	+25	625
50	+15	225
30	-5	25
40	+5	25
20	-15	225
210		1750

3-3 Measures of Variation - Sample Variance

The unbiased estimator of the population variance or the sample variance is a statistic whose value approximates the expected value of a population variance. It is denoted by s^2, where

$$s^2 = \frac{\sum (X - \bar{X})^2}{n-1}, \quad \textit{and}$$

$$\bar{X} = \text{sample mean}$$

$$n = \text{sample size}$$

3-59

3-3 Measures of Variation - Sample Standard Deviation

The sample standard deviation is the square root of the sample variance.

$$s = \sqrt{s^2} = \sqrt{\frac{\sum (X - \bar{X})^2}{n-1}}.$$

3-3 Shortcut Formula for the Sample Variance and the Standard Deviation

$$s^2 = \frac{\sum X^2 - (\sum X)^2 / n}{n-1}$$

$$s = \sqrt{\frac{\sum X^2 - (\sum X)^2 / n}{n-1}}$$

3-3 Sample Variance - Example

- Find the variance and standard deviation for the following sample: 16, 19, 15, 15, 14.
- $\Sigma X = 16 + 19 + 15 + 15 + 14 = 79$.
- $\Sigma X^2 = 16^2 + 19^2 + 15^2 + 15^2 + 14^2 = 1263$.

3-3 Sample Variance - Example

$$s^2 = \frac{\sum X^2 - (\sum X)^2 / n}{n-1}$$

$$= \frac{1263 - (79)^2 / 5}{4} = 3.7$$

$$s = \sqrt{3.7} = 1.9.$$

3-3 Sample Variance for Grouped and Ungrouped Data

- For grouped data, use the class midpoints for the observed value in the different classes.
- For ungrouped data, use the same formula (see next slide) with the class midpoints, X_m, replaced with the actual observed X value.

3-3 Sample Variance for Grouped and Ungrouped Data

The sample variance for grouped data:

$$s^2 = \frac{\sum f \cdot X_m^2 - [(\sum f \cdot X_m)^2 / n]}{n-1}.$$

For ungrouped data, replace X_m with the observe X value.

3-3 Sample Variance for Grouped Data - Example

X	f	$f \cdot X$	$f \cdot X^2$
5	2	10	50
6	3	18	108
7	8	56	392
8	1	8	64
9	6	54	486
10	4	40	400
	$n = 24$	$\Sigma f \cdot X = 186$	$\Sigma f \cdot X^2 = 1500$

3-66

3-3 Sample Variance for Ungrouped Data - Example

The sample variance and standard deviation:

$$s^2 = \frac{\sum f \cdot X^2 - [(\sum f \cdot X)^2 / n]}{n-1}$$

$$= \frac{1500 - [(186)^2 / 24]}{23} = 2.54.$$

$$s = \sqrt{2.54} = 1.6.$$

3-3 Coefficient of Variation

- The **coefficient of variation** is defined to be the standard deviation divided by the mean. The result is expressed as a percentage.

$$\boxed{CVar = \frac{s}{\overline{X}} \cdot 100\% \quad or \quad CVar = \frac{\sigma}{\mu} \cdot 100\%}$$

Chebyshev's Theorem

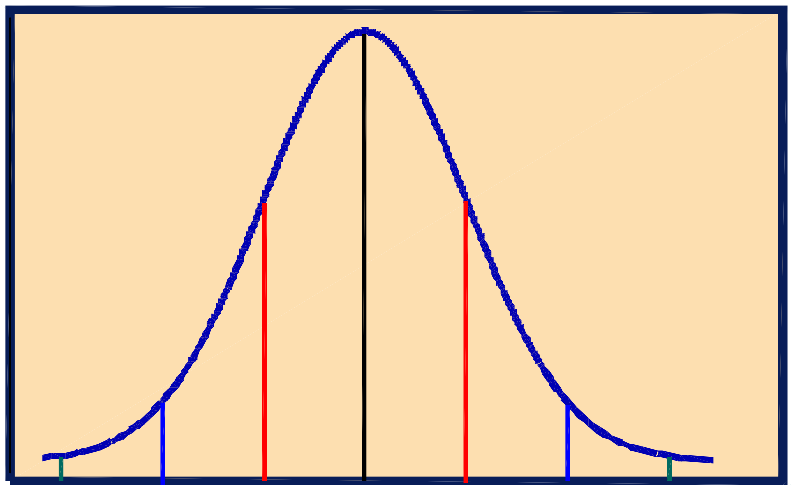
- The proportion of values from a data set that will fall within *k* standard deviations of the mean will be at least $1 - 1/k^2$, where *k* is any number greater than 1.
- For $k = 2$, 75% of the values will lie within 2 standard deviations of the mean. For $k = 3$, approximately 89% will lie within 3 standard deviations.

The Empirical (Normal) Rule

- For any bell shaped distribution:
- Approximately 68% of the data values will fall within one standard deviation of the mean.
- Approximately 95% will fall within two standard deviations of the mean.
- Approximately 99.7% will fall within three standard deviations of the mean.

The Empirical (Normal) Rule

$\mu \pm 1\sigma$ –– 68% $\mu \pm 2\sigma$ -- 95% $\mu \pm 3\sigma$ –– 99.7%



$\mu - 3\sigma$ $\mu - 2\sigma$ $\mu - 1\sigma$ μ $\mu + 1\sigma$ $\mu + 2\sigma$ $\mu + 3\sigma$

3-4 Measures of Position — *z* score

- The **standard score** or ***z* score** for a value is obtained by subtracting the mean from the value and dividing the result by the standard deviation.
- The symbol z is used for the ***z* score**.

3-4 Measures of Position — z-score

- The *z* score represents the number of standard deviations a data value falls above or below the mean.

For samples:

$$z = \frac{X - \overline{X}}{s}.$$

For populations:

$$z = \frac{X - \mu}{\sigma}.$$

3-4 z-score - Example

- A student scored 65 on a statistics exam that had a mean of 50 and a standard deviation of 10. Compute the *z*-score.
- $z = (65 - 50)/10 = 1.5$.
- That is, the score of 65 is 1.5 standard deviations above the mean.
- Above - since the *z*-score is positive.

3-4 Measures of Position - Percentiles

- **Percentiles divide the distribution into 100 groups.**
- **The P_k percentile is defined to be that numerical value such that at most k% of the values are smaller than P_k and at most $(100 - k)$% are larger than P_k in an ordered data set.**

3-4 Percentile Formula

- **The percentile corresponding to a given value (*X*) is computed by using the formula:**

$$Percentile = \frac{\text{number of values below } X + 0.5}{\text{total number of values}} \cdot 100\%$$

3-4 Percentiles - Example

- A teacher gives a 20-point test to 10 students. Find the percentile rank of a score of 12. Scores: 18, 15, 12, 6, 8, 2, 3, 5, 20, 10.
- Ordered set: 2, 3, 5, 6, 8, 10, 12, 15, 18, 20.
- Percentile = [(6 + 0.5)/10](100%) = 65th percentile. Student did better than 65% of the class.

3-4 Percentiles - Finding the value Corresponding to a Given Percentile

- **Procedure:** Let p be the percentile and n the sample size.
- **Step 1:** Arrange the data in order.
- **Step 2:** Compute $c = (n \cdot p)/100$.
- **Step 3:** If c is not a whole number, round up to the next whole number. If c is a whole number, use the value halfway between c and $c+1$.

3-78

3-4 Percentiles - Finding the value Corresponding to a Given Percentile

- **Step 4:** The value of *c* is the position value of the required percentile.
- **Example:** Find the value of the 25th percentile for the following data set: 2, 3, 5, 6, 8, 10, 12, 15, 18, 20.
- *Note:* the data set is already ordered.
- $n = 10$, $p = 25$, so $c = (10 \cdot 25)/100 = 2.5$. Hence round up to $c = 3$.

3-4 Percentiles - Finding the value Corresponding to a Given Percentile

- Thus, the value of the 25th percentile is the value $X = 5$.
- Find the 80th percentile.
- $c = (10 \cdot 80)/100 = 8$. Thus the value of the 80th percentile is the average of the 8th and 9th values. Thus, the 80th percentile for the data set is $(15 + 18)/2 = 16.5$.

3-4 Special Percentiles - Deciles and Quartiles

- **Deciles** divide the data set into 10 groups.
- Deciles are denoted by D_1, D_2, …, D_9 with the corresponding percentiles being P_{10}, P_{20}, …, P_{90}
- **Quartiles** divide the data set into 4 groups.

3-4 Special Percentiles - Deciles and Quartiles

- Quartiles are denoted by Q_1, Q_2, and Q_3 with the corresponding percentiles being P_{25}, P_{50}, and P_{75}.
- The median is the same as P_{50} or Q_2.

3-4 Outliers and the Interquartile Range (IQR)

- An outlier is an extremely high or an extremely low data value when compared with the rest of the data values.
- The Interquartile Range, $IQR = Q_3 - Q_1$.

3-4 Outliers and the Interquartile Range (IQR)

- **To determine whether a data value can be considered as an outlier:**
- **Step 1:** Compute Q_1 and Q_3.
- **Step 2:** Find the IQR = $Q_3 - Q_1$.
- **Step 3:** Compute (1.5)(IQR).
- **Step 4:** Compute $Q_1 - (1.5)(\text{IQR})$ and $Q_3 + (1.5)(\text{IQR})$.

3-4 Outliers and the Interquartile Range (IQR)

- **To determine whether a data value can be considered as an outlier:**
- **Step 5: Compare the data value (say X) with Q_1 – (1.5)(IQR) and Q_3 + (1.5)(IQR).**
- **If $X < Q_1 - (1.5)(IQR)$ or if $X > Q_3 + (1.5)(IQR)$, then X is considered an outlier.**

3-85

3-4 Outliers and the Interquartile Range (IQR) - Example

- **Given the data set 5, 6, 12, 13, 15, 18, 22, 50, can the value of 50 be considered as an outlier?**
- **$Q_1 = 9$, $Q_3 = 20$, IQR = 11. *Verify.***
- **(1.5)(IQR) = (1.5)(11) = 16.5.**
- **9 – 16.5 = – 7.5 and 20 + 16.5 = 36.5.**
- **The value of 50 is outside the range – 7.5 to 36.5, hence 50 is an outlier.**

3-5 Exploratory Data Analysis - Stem and Leaf Plot

- A **stem and leaf plot** is a data plot that uses part of a data value as the stem and part of the data value as the leaf to form groups or classes.

3-5 Exploratory Data Analysis - Stem and Leaf Plot - Example

- At an outpatient testing center, a sample of 20 days showed the following number of cardiograms done each day: 25, 31, 20, 32, 13, 14, 43, 02, 57, 23, 36, 32, 33, 32, 44, 32, 52, 44, 51, 45. Construct a stem and leaf plot for the data.

3-88

3-5 Exploratory Data Analysis - Stem and Leaf Plot - Example

Leading Digit (Stem)	Trailing Digit (Leaf)
0	2
1	3 4
2	0 3 5
3	1 2 2 2 2 3 6
4	3 4 4 5
5	1 2 7

3-5 Exploratory Data Analysis - Box Plot

- When the data set contains a small number of values, a box plot is used to graphically represent the data set. These plots involve five values: the minimum value, the lower hinge, the median, the upper hinge, and the maximum value.

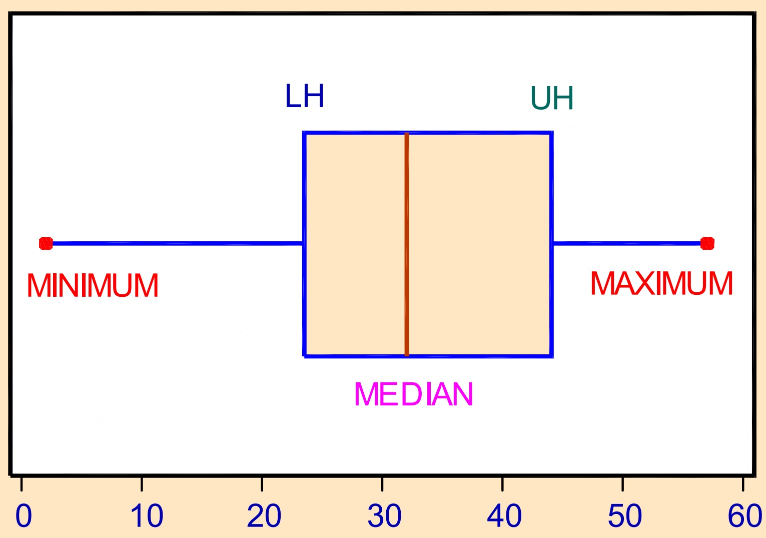
3-5 Exploratory Data Analysis - Box Plot

- The lower hinge is the median of all values less than or equal to the median when the data set has an odd number of values, or as the median of all values less than the median when the data set has an even number of values. The symbol for the lower hinge is LH.

3-5 Exploratory Data Analysis - Box Plot

- The upper hinge is the median of all values greater than or equal to the median when the data set has an odd number of values, or as the median of all values greater than the median when the data set has an even number of values. The symbol for the upper hinge is UH.

3-5 Exploratory Data Analysis - Box Plot - Example (Cardiograms data)



Information Obtained from a Box Plot

- If the median is near the center of the box, the distribution is approximately symmetric.
- If the median falls to the left of the center of the box, the distribution is positively skewed.
- If the median falls to the right of the center of the box, the distribution is negatively skewed.

3-94

Information Obtained from a Box Plot

- **If the lines are about the same length, the distribution is approximately symmetric.**
- **If the right line is larger than the left line, the distribution is positively skewed.**
- **If the left line is larger than the right line, the distribution is negatively skewed.**