Chapter 02

Descriptive Statistics (Measures)

Biostatistics For the Health Sciences

Dr. Moustafa Omar Ahmed Abu-Shawiesh
Professor of Statistics

2.1 Introduction

This chapter shows the statistical methods that can be used to summarize (describe) a data set. We will learn how to calculate and interpret the descriptive measures [that is, the value(s) which are used to describe a data set] such as measures of center, measures of variation, and measures of position.

Descriptive Statistics

Consists of methods for organizing, displaying, and describing data by using tables, graphs, and summary measures.

Notation

Features of good numeric or graphic form of data summarization:

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

2.2 Measures of Location (Mean, Median, and Mode)

- It is easy to lose track of the overall picture when there are too many sample points (observations).
- Data summarization is important before any inferences (conclusions) can be made about the population from which the sample points have been obtained.

Definition: Measure of Location

It is a type of measure useful for data summarization that defines the center or middle value (most typical value) of the data set (sample).

Notation

The most commonly used measures of location are: **Mean**, **Median**, and **Mode**.

The Arithmetic Mean

Consider a random sample of n data points (Sample Size) $x_1, x_2, \ldots, x_n$ drawn from some population of size N (Population Size), then the mean or average or sample mean or arithmetic mean can be defined as follows:

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

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

Notation

The arithmetic mean (or mean or average or sample mean) is usually denoted by x-bar ($\bar{x}$). Sigma ($\sum$) is a summation sign, that is, $\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n$.

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

Example (Infants Birthweights)

Find the value of the arithmetic mean ($\bar{x}$) for the sample of birthweights given below 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

Solution

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{i=1}^{20} x_i}{20} = \frac{3265 + 3260 + \ldots + 2834}{20} = \frac{63338}{20} = 3166.9 \text{ gram or } 3.1669 \text{ Kilogram}$$

The Median

An alternative measure of location, perhaps second in popularity to the arithmetic mean, is the median or, more precisely, the sample median. Suppose there are n observations in the random sample, say $x_1, x_2, \ldots, x_n$, then if these observations are ordered (ascending order) from smallest value (minimum) to the largest value (maximum), that is $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, then the sample median can be 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

For example, for samples of size n = 7, the fourth largest point is the central point in the sense that 3 points are smaller than it and 3 points are larger. Thus, for samples of size n = 8 the fourth and fifth largest points would be averaged to obtain the median, because neither is the central point. The main strength of the sample median is that it is insensitive to very large or very small values (Outliers).

Example (Infants Birthweights)

Find the value of the sample median for the sample of birthweights given below 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

Solution

First, arrange the sample in an ascending order as follows:

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

Second, because n = 20 is an even number, then the sample median can be calculated as follows:

$$\text{Sample median} = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} = \frac{x_{\left(\frac{20}{2}\right)} + x_{\left(\frac{20}{2}+1\right)}}{2} = \frac{x_{(10)} + x_{(11)}}{2} = \frac{3245 + 3248}{2}$$

$$= \frac{6493}{2} = 3246.5 \text{ gram or } 3.2465 \text{ Kilogram}$$

Example (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:

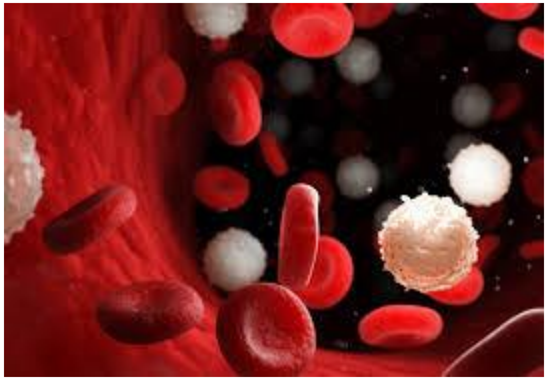


TABLE 2.3 Sample of admission white-blood counts (× 1000) for all patients entering a hospital in Allentown, Pennsylvania, on a given day

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

Compute the sample median for white-blood counts?

Solution

First, arrange the sample in an ascending order as follows: 3, 5, 7, 8, 8, 9, 10, 12, 35.

Second, because n = 9 is an odd number, then the sample median can be calculated as follows:

Sample median $= x_{\left(\frac{n+1}{2}\right)} = x_{\left(\frac{9+1}{2}\right)} = x_{\left(\frac{10}{2}\right)} = x_{(5)} = 8$ (or 8000 on the original scale)

Notation: Suppose that 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

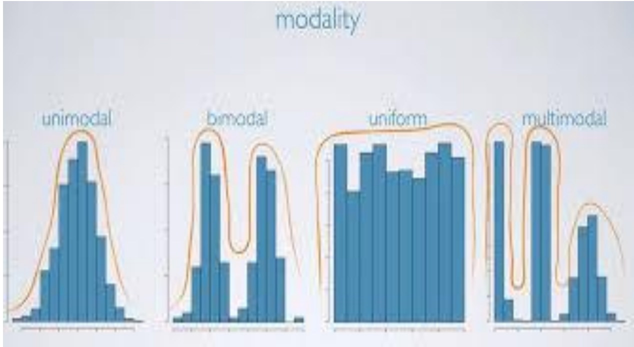
The mode is another widely used measure of location.

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

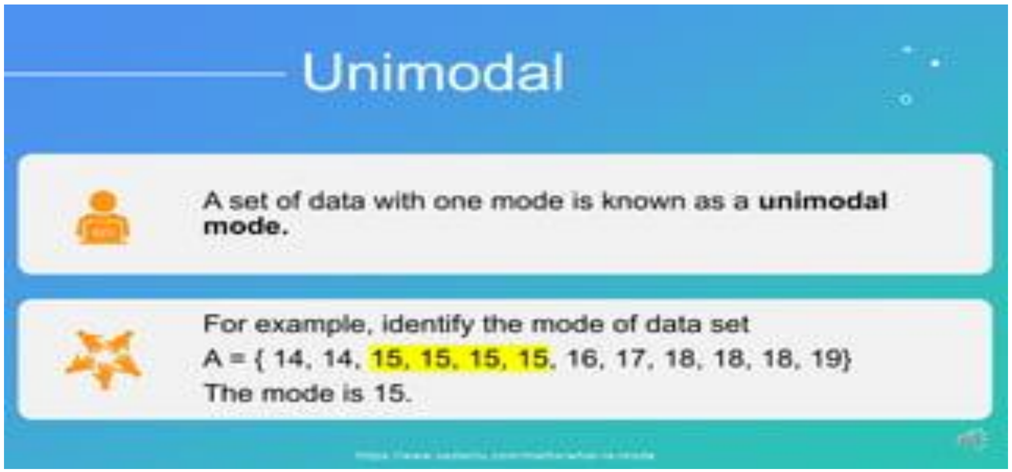
Notation

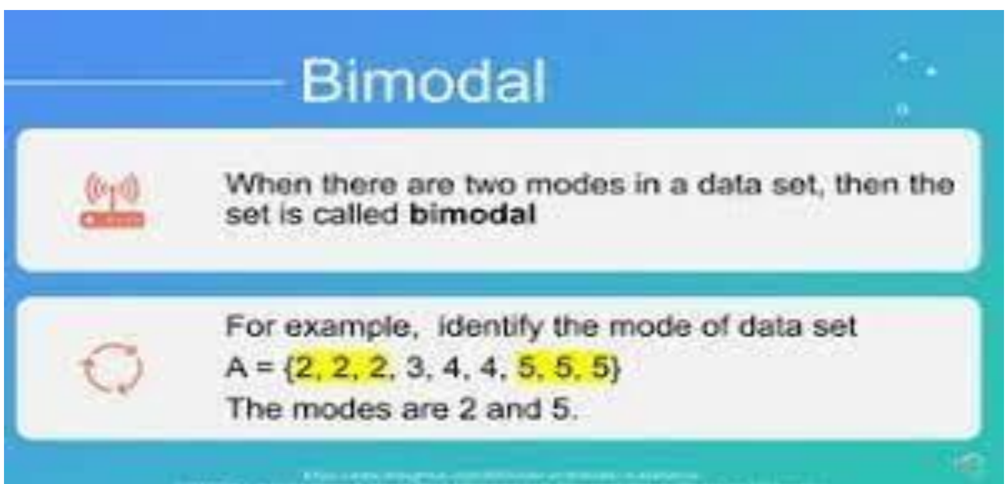
Any data set selected from a given distribution may have no modes or one mode or more than one mode, and therefore according to the number of modes, we can divided the distribution shape into:

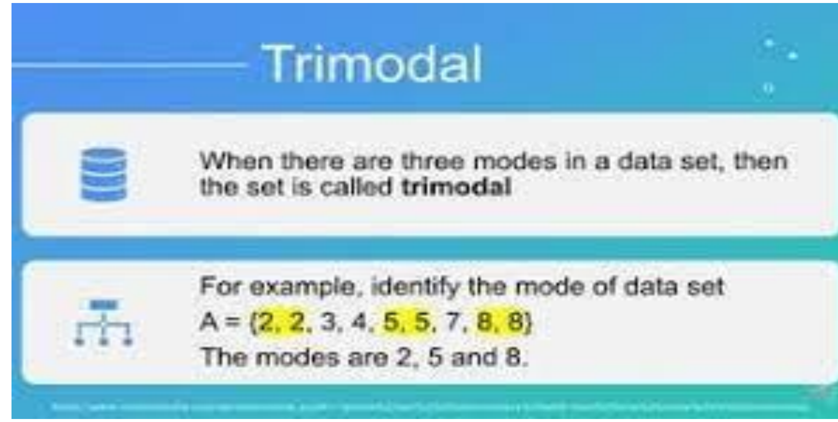
- One mode = Unimodal Distribution.
- Two modes = Bimodal Distribution.
- Three modes = Trimodal distribution.
- More than three modes = Multimodal distribution.

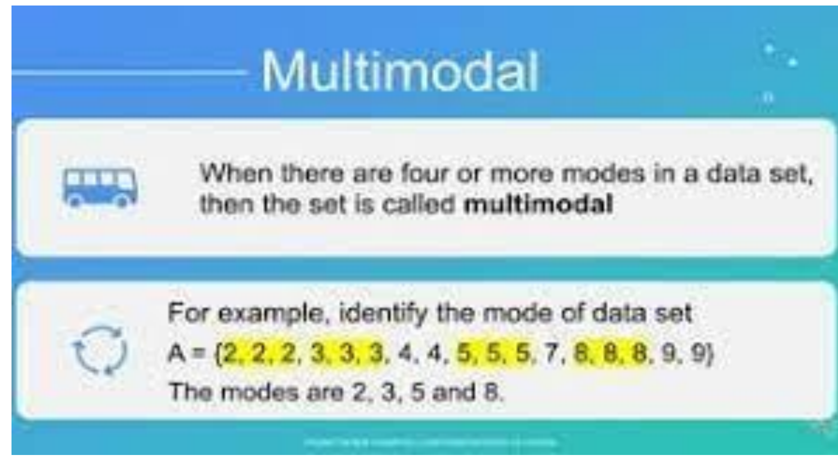


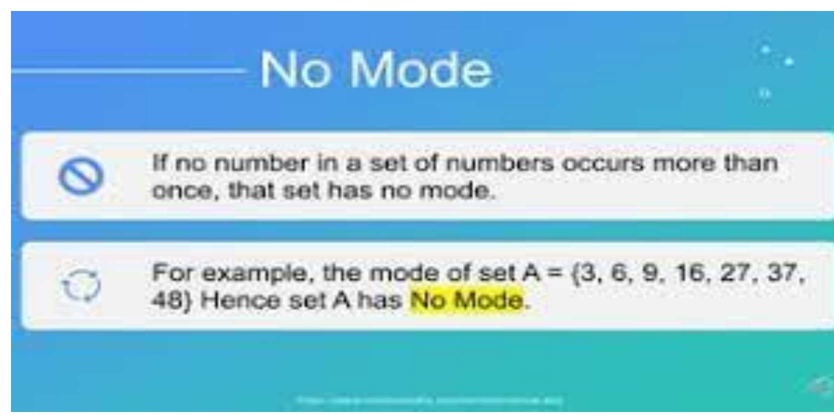
Important: Some distributions have more than one mode. In fact, one useful method of classifying distributions is by the number of modes present. A distribution with one mode is called unimodal; two modes, bimodal; three modes, trimodal; and so forth.











Example (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. What is the value of the mode?

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

Solution: The mode is 28 because it is the most frequently occurring value.

Example: 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.

Example: Compute the mode of the distribution in Table 2.1?

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

Notation: This example 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.

Example

A survey on the Ministry of Health showed the following distribution for the number of tablets sold in May 2021 for five types of medications used to treat systolic blood pressure:

Medicine Name	Number of Tablets Sold
Almor	632
Lasix	1425
Aldacton	878
Indicardin	95
Diovan	471

Find the mode?

Solution: Since the category with the highest frequency is Lasix, then the mode for the number of tablets sold in May 2021 for the five types of medications used to treat systolic blood pressure is the Lasix drug.

Example

The blood types for 30 people who agreed to participate in a medical study were as follows.

O A A O A AB O B A O
A O A B O O O AB A A
A B O A A O O B O O

Find the mode of the blood types.

Solution:

Blood Type	Frequency
A	11
B	4
AB	2
O	13

Mode = O

Comparison of the Arithmetic Mean and the Sample Median

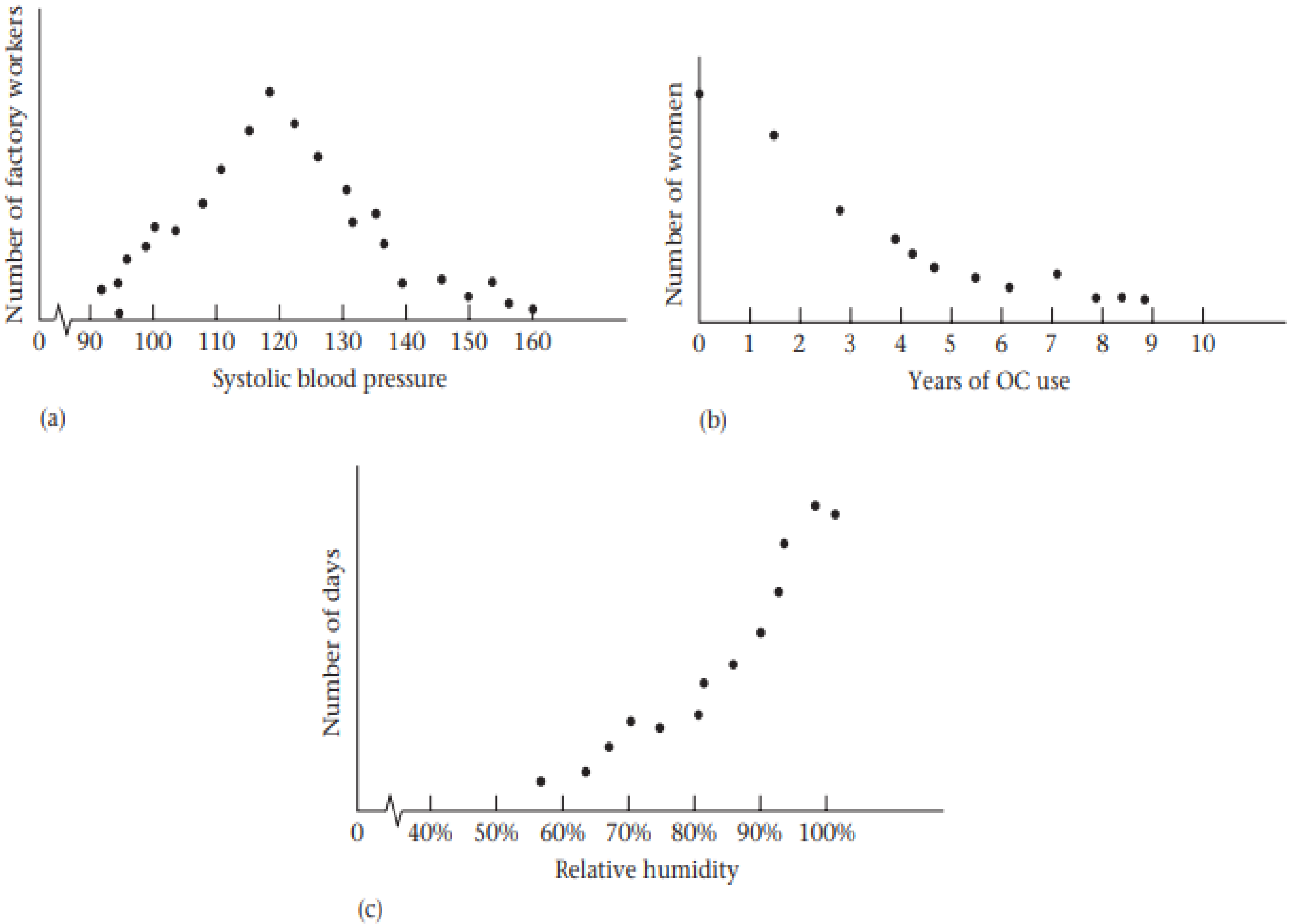
In many samples, the relationship between the arithmetic mean and the sample median can be used to assess the symmetry of a distribution as follows:

- If a distribution is symmetric, then the arithmetic mean is approximately the same as the sample median.
  An example of a distribution that is expected to be roughly symmetric is the distribution of systolic blood-pressure measurements taken on all 30-to-39 year old factory workers in a given workplace (Figure 2.3a).

- If a distribution is positively skewed (skewed to the right), then the arithmetic mean tends to be larger than the sample median.
  An example of a positively skewed distribution is that of the number of years of oral contraceptive (OC) use among a group of women ages 20 to 29 years (Figure 2.3b).

- If a distribution is negatively skewed (skewed to the left), then the arithmetic mean tends to be smaller than the sample median.
  An example of a negatively skewed distribution is that of relative humidities observed in a humid climate at the same time of day over a number of days. In this case, most humidities are at or close to 100%, with a few very low humidities on dry days (Figure 2.3c).

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



2.3 Some Properties of the Arithmetic Mean

Property (1): Consider a random sample of size (n); $x_1, x_2, \ldots, x_n$; which will be referred to as the original sample. To create a translated sample, add a constant c to each data point, to get $x_1 + c$, $x_2 + c$, … , $x_n + c$. Let $y_i = x_i + c, i = 1, \ldots, n$. Suppose we want to compute the arithmetic mean of the translated sample, we can show that the following relationship (Equation 2.1) holds:

EQUATION 2.1

If $\quad y_i = x_i + c, \quad i = 1, \ldots, n$

then $\quad \bar{y} = \bar{x} + c$

Example (Infants Birthweights)

Express the mean birthweight in grams for the data in Table 2.1 if the following equation $y_i = x_i - 250, i = 1, \ldots, 20$ is used?

Solution

The mean for the original sample is:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{i=1}^{20} x_i}{20} = 3166.9 \text{ gram}$$

Then the new mean for the translated sample is given by:

$$\bar{y} = \bar{x} - 250 = 3166.9 - 250 = 2916.9 \text{ gram or } 2.9169 \text{ kilogram.}$$

Question: What happens to the arithmetic mean if the units or scale being worked with changes? To answer this question, we will consider the following property:

Property (2): Consider a random sample of size (n); $x_1, x_2, \ldots, x_n$; which will be referred to as the original sample. To create a rescaled sample, multiply a constant c by each data point, to get cx_1 , cx_2 , … , cx_n. Let $y_i = cx_i$, $i = 1, \ldots, n$. Suppose we want to compute the arithmetic mean of the rescaled sample, we can show that the following relationship (Equation 2.2) holds:

EQUATION 2.2

If	$y_i = cx_i,\ i = 1, \ldots, n$
then	$\bar{y} = c\bar{x}$

Example (Infants Birthweights)

Express the mean birthweight for the data in Table 2.1 in ounces rather than grams?

Solution

We know that 1 oz = 28.35 gram and that $\bar{x}$ = 3166.9 gram. Thus, if the data were expressed in terms of ounces, we have:

$$c = \frac{1}{28.35}$$

$$\text{and } \bar{y} = c\bar{x} = \frac{1}{28.35}(3166.9) = 111.71 \text{ oz}$$

Property (3): Linear Transformation

Sometimes we want to change both the origin and the scale of the original data set $x_1, x_2, \ldots, x_n$ at the same time. To do this, we apply Equations 2.1 and 2.2 as follows:

EQUATION 2.3

Let $x_1, \ldots, x_n$ be the original sample of data and let $y_i = c_1 x_i + c_2$, $i = 1, \ldots, n$ 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 $\quad y_i = c_1 x_i + c_2, \quad i = 1, \ldots, n$

then $\quad \bar{y} = c_1 \bar{x} + c_2$

Example (Temperature)

If we have a random sample of temperatures in °C with an arithmetic mean of $\bar{x}$ = 11.75 °C, then what is the arithmetic mean in °F?

Solution

Let y_i denote the °F temperature that corresponds to a °C temperature of x_i. The required transformation to convert the data to °F would be as follows:

$$y_i = \frac{9}{5} x_i + 32, \quad i = 1, \ldots, n$$

Implies that $\bar{y} = \frac{9}{5}\bar{x} + 32$

so the arithmetic mean would be $\bar{y} = \frac{9}{5}(11.75) + 32 = 53.15°\text{F}$.

2.4 Measures of Variation (Dispersion or Spread)

A measure of variation (dispersion) gives information regarding the amount of variability present in a data set.

Notations

1. If all the values are the same → no dispersion (no variation).
2. If all the values are different → there is a dispersion.
3. If the values close to each other → dispersion is small.
4. If the values are widely scattered → dispersion is large.

Most Commonly Used Measures of Variation (Dispersion)

Several different measures can be used to describe the variability of a sample. The most commonly used measures of variation are:

1. Range (R) 2. Variance 3. Standard Deviation and 4. Coefficient of Variation (CV).

The Range

The range (R) is the simplest measure of variation and can be defined as follows:

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

$$R = Max - Min = X_{(n)} - X_{(1)}$$

Example

The age in years for a random sample of 10 patients selected from the emergency room of KAUH on a Friday night at February 2021 are given as follows: 43, 66, 61, 64, 65, 38, 59, 57, 57, 50. Find the value of the range for ages?

Solution

Maximum Age = X(n) = 66 year

Minimum Age = X(1) = 38 year

then: R = 66 – 38 = 28 year.

Note that: The sample mean ($\overline{x}$) is equal to 56 year.

Notation: 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 and therefore the range is considered as a poor measure of variation.

The Variance and Standard Deviation

In order to have a more meaningful statistic to measure the variability , we use measures called the variance and standard deviation.

Let x_1, x_2, ..., x_n be a random sample of size n observations (raw data) selected from a population of size N, then the sample variance denoted by s^2 can be calculated and defined as follows:

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

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

Another commonly used measure of spread is the sample standard deviation (s) defined as follows:

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

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

where the sample mean ($\bar{x}$) is given as follows:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Question: Why is it necessary to take the square root for variance?

Answer: The reason is that since the distances were squared, the units of the resultant numbers are the squares of the units of the original raw data, then finding the square root of the variance puts the standard deviation in the same units of the raw data.

Example

Calculate the sample variance (s^2) and sample standard deviation (s) of the birthweight data in Table 2.1 in grams:

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

Solution

- The value of the arithmetic mean (sample mean ($\bar{x}$)) is calculated as follows:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{i=1}^{20} x_i}{20} = \frac{3265 + 3260 + \dots + 2834}{20} = \frac{63338}{20} = 3166.9 \text{ gram or } 3.1669 \text{ Kilogram}$$

- The value of the sample variance (s^2) is calculated as follows:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{(3265 - 3166.9)^2 + \dots + (2834 - 3166.9)^2}{19}$$

$$= \frac{9623.61 + \dots + 110822.41}{19} = \frac{3768147.8}{19} = 198323.6 \text{ gram}^2$$

- The value of the sample standard deviation (s) is calculated as follows:

$$s = \sqrt{s^2} = \sqrt{198323.6} = 445.3 \text{ gram}$$

Notation: the arithmetic mean and the standard deviation are in the same units, whereas the arithmetic mean and the variance are not. Also, the mean and standard deviation are the most widely used measures of location and variation in the literature. One of the main reasons for this is that the normal (or bell-shaped) distribution is defined explicitly in terms of these two parameters (measures), and this distribution has wide applicability in many biological and medical settings. The normal distribution is discussed extensively in Chapter 5.

2.5 Some Properties of the Variance and Standard Deviation

Property (1)

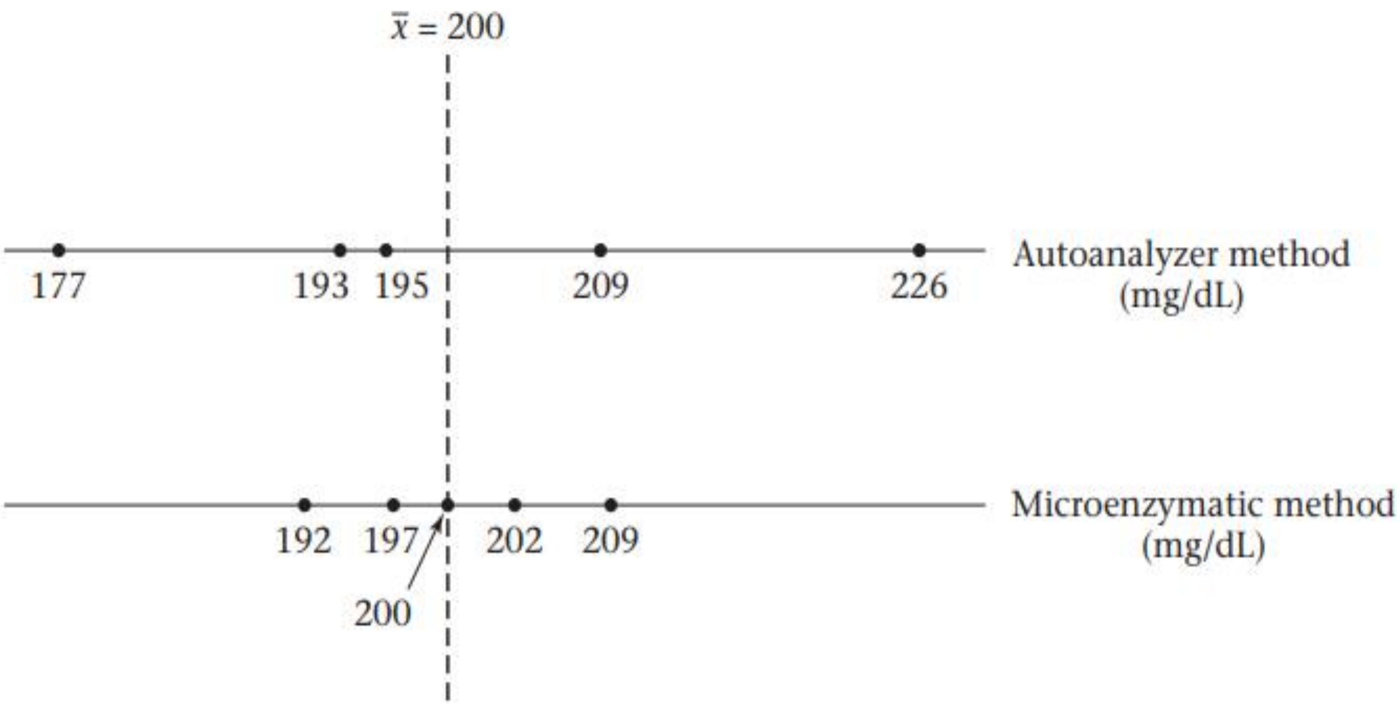
EQUATION 2.4 The sum of the deviations of the individual observations of a sample about the sample mean is always zero.

$$d = \sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

Example

Compute the sum of the deviations about the mean for the Autoanalyzer method data in Figure 2.4.

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



Solution

For the Autoanalyzer-method data,

$$d = (177 - 200) + (193 - 200) + (195 - 200) + (209 - 200) + (226 - 200)$$
$$= -23 - 7 - 5 + 9 + 26 = 0$$

Question: How are the variance and standard deviation affected by a change in origin or a change in the units being worked with?

Property (2)

EQUATION 2.5

Suppose there are two samples

$$x_1, \ldots, x_n \quad \text{and} \quad y_1, \ldots, y_n$$

where $y_i = x_i + c, \quad i = 1, \ldots, n$

If the respective sample variances of the two samples are denoted by

$$s_x^2 \text{ and } s_y^2$$

then $s_y^2 = s_x^2$

Property (3)

EQUATION 2.6

Suppose there are two samples

$$x_1, \ldots, x_n \quad \text{and} \quad y_1, \ldots, y_n$$

where $y_i = cx_i, \quad i = 1, \ldots, n, \quad c > 0$

Then $s_y^2 = c^2 s_x^2 \quad s_y = cs_x$

Example

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:

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

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

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

Quantiles (Percentiles)

It is another approach that addresses some of the shortcomings of the range (R) in quantifying the spread. The pth percentile is the value Vp such that p percent of the sample points are less than or equal to Vp.

Now, depending on whether or not $(np/100)$ is an integer, the definition for the pth percentile can be given as follows:

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.**

Notations

(1) To compute percentiles, the sample points must be ordered.

(2) The median, being the $50th$ percentile, is a special case of a quantile.

(3) Frequently used percentiles are quartiles (25^{th}, 50^{th}, and 75^{th} percentiles).

Example

Compute the $20th$ percentile for the white-blood-count data in Table 2.3:

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

(1) Arrange the sample data points in an ascending order as follows:

3, 5, 7, 8, 8, 9, 10, 12, 35

(2) Because $np/100 = 9 \times 0.2 = 1.8$ is not an integer, and therefore we round up the value 1.8 to the next integer, then the $20th$ percentile is defined by the $(1 + 1)th$ largest value = second largest value = $x_{(2)}$= 5000.

Example

Compute the $10th$ and $90th$ percentiles for the birthweight data 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

Solution

(1) Arrange the sample data points in an ascending order as follows:

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

(2) Because 20 × 0.1 = 2 and 20 × 0.9 = 18 are integers, the $10th$ and $90th$ percentiles are defined by:

$10th\ percentile$

Average of second and third largest values = $\frac{x_{(2)} + x_{(3)}}{2}$ = (2581 + 2759)/2 = 2670 g.

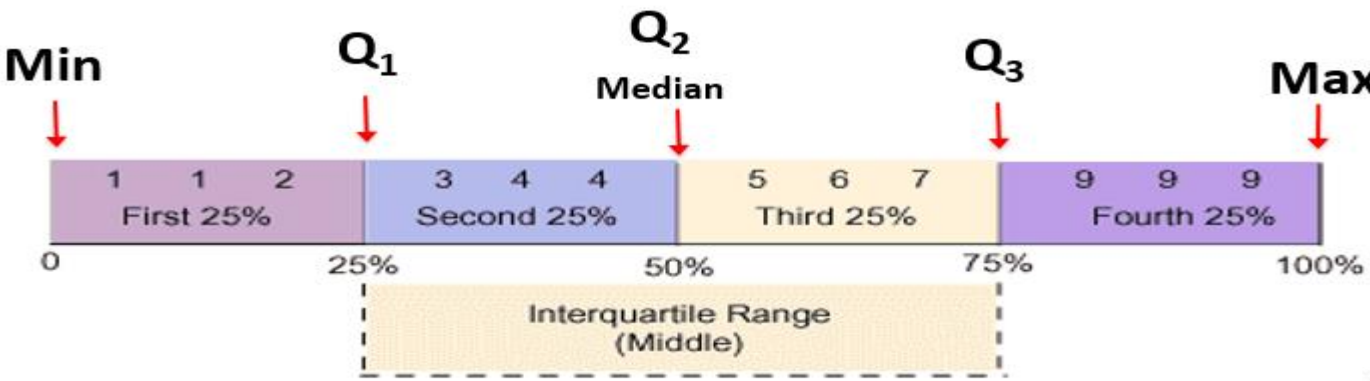
$90th\ percentile$

Average of 18th and 19th largest values = $\frac{x_{(18)} + x_{(19)}}{2}$ = (3609 + 3649)/2 = 3629 g.

Conclusion: We would estimate that 80% of birthweights will fall between 2670 g and 3629 g.

Interquartile Range (IQR)

Quartiles are the values of observations in a data set, when arranged in an ordered sequence, can divided the data set into four equal parts, or quarters, using three quartiles namely Q1, Q2 and Q3 each representing a fourth of the population being sampled.



Notation

First (lower) Quartile (Q1) = 25th Percentile.

Second Quartile (Q2) = Median = 50th Percentile.

Third (upper) Quartile (Q3) = 75th Percentile.

Definition: The interquartile range (IQR) is a robust measure of variation that is based on the quartiles. The IQR is defined as the range of the middle 50% of the observations in the data set. It is the difference between the third quartile (Q3) and the first quartile (Q1) and found by using the following formula:

$$IQR = Q3 - Q1$$

Outliers (Outlying Values)

The IQR can help identify possible outlying values—that is:

Values that seem inconsistent with the rest of the points in the random sample.

In this context, outlying values are defined as follows:

DEFINITION 2.11 An **outlying value** is a value x such that either

(1) $x >$ upper quartile + 1.5 × (upper quartile – lower quartile) or

(2) $x <$ lower quartile – 1.5 × (upper quartile – lower quartile)

DEFINITION 2.12 An **extreme outlying value** is a value x such that either

(1) $x >$ upper quartile + 3.0 × (upper quartile – lower quartile) or

(2) $x <$ lower quartile – 3.0 × (upper quartile – lower quartile)

Example

Using the white-blood-count data in Table 2.3, comment on the presence of outlying values?

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

(1) Arrange the sample data points in an ascending order as follows:

3, 5, 7, 8, 8, 9, 10, 12, 35

(2) The lower and upper quartiles are calculated as follows:

Lower Quartile (Q1)

$np/100 = 9 \times 0.25 = 2.25$ is not an integer, and therefore we round up the value 2.25 to the next integer 3, then the $25th$ percentile is the third largest value = $x_{(3)}$= 7000.

Upper Quartile (Q3)

$np/100 = 9 \times 0.75 = 6.75$ is not an integer, and therefore we round up the value 6.75 to the next integer 7, then the $75th$ percentile is the seventh largest value = $x_{(7)}$= 10000.

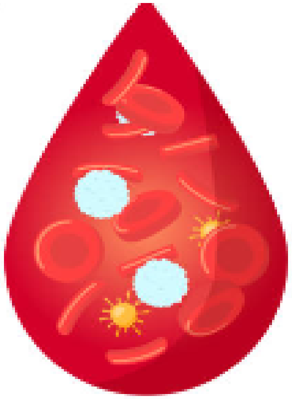
Interquartile Range (IQR): IQR = Q3 – Q1 = 10000 – 7000 = 3000

Outlying Values

Q1 – 1.5 IQR = 7000 – (1.5 x 3000) = 7000 – 4500 = 2500

Q3 + 1.5 IQR = 10000 + (1.5 x 3000) = 10000 + 4500 = 14500

Then x = 35000 is an outlying value because it is > 14500.



Extreme Outlying Values

Q1 – 3 IQR = 7000 – (3 x 3000) = 7000 – 9000 = - 2000

Q3 + 3 IQR = 10000 + (3 x 3000) = 10000 + 9000 = 19000

Then x = 35000 is an extreme outlying value because it is > 19000.

2.6 The Coefficient of Variation (CV)

It is useful to relate the arithmetic mean and the standard deviation to each other, a special measure, the coefficient of variation (CV), is often used for this purpose and can be defined as follows:

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

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

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. The CV is most useful in comparing the variability of several different samples, each with different arithmetic means. It is a relative measure of variability.

Example

Two health clubs A and B from Jordan show the following results about the number of workers and the monthly wages in JD paid to them:

Health Club	No. of Workers	Sample Mean	Sample Standard Deviation
A	50	250 JD	9 JD
B	60	350 JD	10 JD

In which health club, A or B, is there greater variability in individual wages?

Solution

The coefficient of variation (CV) allows us to make a relative comparison of the variability as follows:

$$CV_{Club\ A} = \frac{S}{\bar{X}} * 100\% = \frac{9}{250} * 100\% = 0.036 * 100\% = 3.60\%$$

$$CV_{Club\ B} = \frac{S}{\bar{X}} * 100\% = \frac{10}{350} * 100\% = 0.0286 * 100\% = 2.86\%$$

Conclusion

Since the value of CV for wages of workers in club A is greater than the value of CV for wages of workers in club B, then club A has more variability which means that wages paid by club B for workers is better and more consists from wages paid for them by club A.

2.8 Graphic Methods

In Sections 2.1 through 2.6 we concentrated on methods for describing data in numeric and tabular form. In this section, these techniques are supplemented by presenting certain commonly used graphic methods for displaying data. The purpose of using graphic displays is to give a quick overall impression of data, which is sometimes difficult to obtain with numeric measures.

Bar Graphs (Bar Charts)

The bar graph is one of the most widely used methods for displaying data. It is a graph showing the differences in frequencies or percentages among categories of a data (ungrouped or grouped). A bar graph can be constructed as follows:

(1) The data are divided into a number of groups (categories).

(2) For each group (category), a rectangle (bar) is constructed with a base of a constant width on the x-axis and a height proportional to the frequency or percentage within that group (category) on the y-axis.

(3) The rectangles (bars) are generally not contiguous (separated) and are equally spaced from each other to emphasize the fact that each bar is a separate category (group).

Example

The patients staying at King Abdullah University Hospital (KAUH) were asked to rate the quality of their stay in the hospital as being *excellent*, *above average*, *average*, *below average*, or *poor*. The ratings provided by a random sample of n = 20 patients. The results are given as follows:

Below Average	Above Average	Above Average	Excellent
Above Average	Above Average	Poor	Average
Average	Below Average	Average	Above Average
Above Average	Poor	Above Average	Below Average
Average	Average	Above Average	Above Average

and shown on the given frequency distribution:

Rating (Group) (x)	Frequency (f)
Poor	2
Below Average	3
Average	5
Above Average	9
Excellent	1
Total	20

Draw the bar graph for the variable "quality of the stay" in the KAUH?

Solution

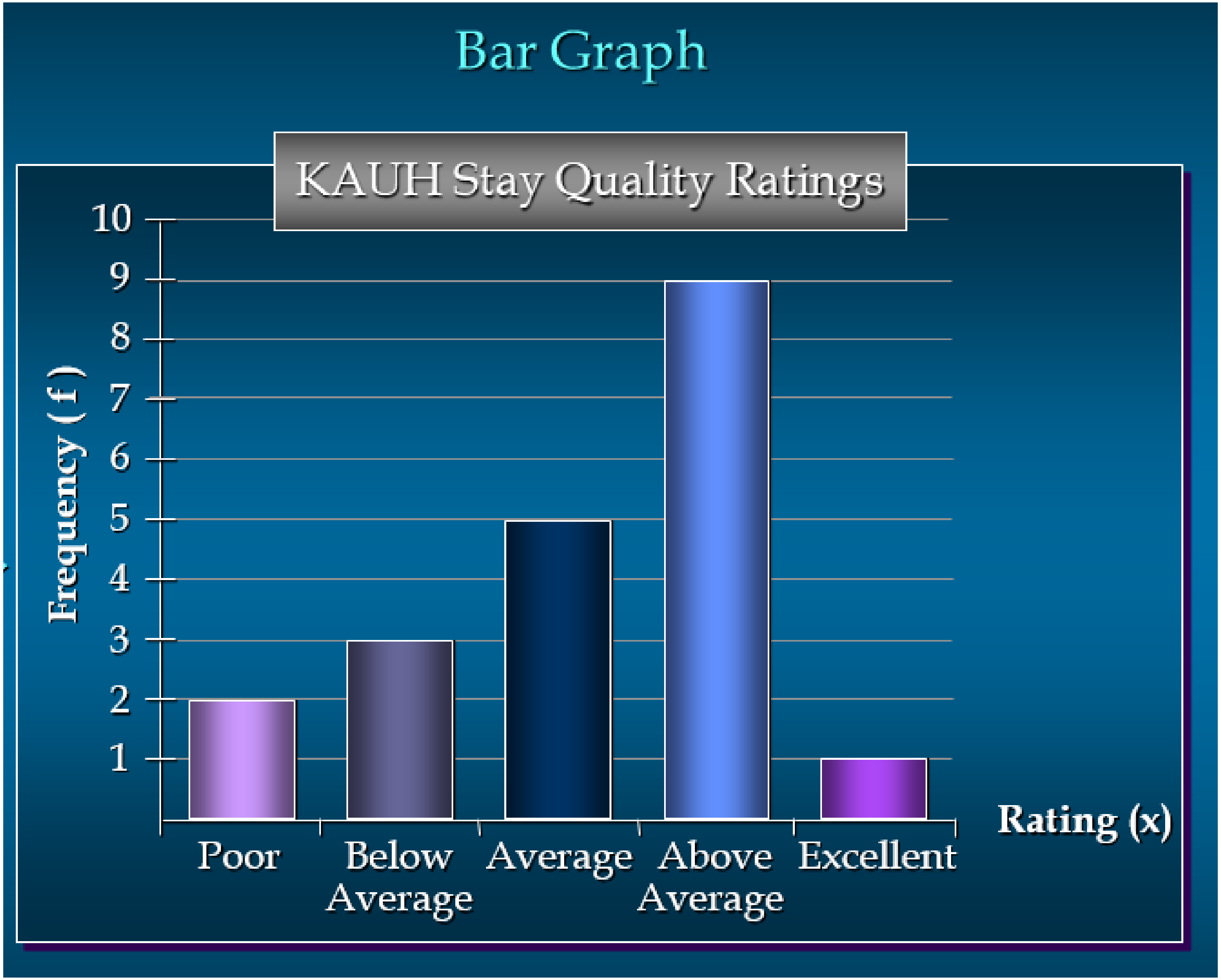


Figure (1): Quality rating for Patients stay in King Abdullah University Hospital (KAUH).

Exercise: The table below gives information about the meals ordered by a random sample of 180 patient from Jordan Hospital on a Friday day in August 2023:

Meal (x)	Frequency (f)
Chicken	54
Beef	75
Pizza	39
Vegetarian	12
Total	180

Draw the bar graph for the variable "Meal" for these information in this hospital?

The Box Plot

In this section, we will discuss the comparison of the arithmetic mean and the median as a method for looking at the skewness of a distribution by a graphic technique known as the box plot. A box plot uses the relationships among the median (Q2), upper quartile (Q3), and lower quartile (Q1) to describe the skewness of a distribution.

Question: How can the median (Q2), upper quartile (Q3), and lower quartile (Q1) be used to judge the symmetry of a distribution?

Answer

(1) If the distribution is symmetric, then the upper and lower quartiles should be approximately equally spaced from the median $(Q2 - Q1 = Q3 - Q2)$.

(2) If the upper quartile is farther from the median than the lower quartile $(Q3 - Q2 > Q2 - Q1)$, then the distribution is positively skewed.

(3) If the lower quartile is farther from the median than the upper quartile $(Q2 - Q1 > Q3 - Q2)$, then the distribution is negatively skewed.

The above relationships are illustrated graphically in a box plot. In addition to displaying the symmetry properties of a sample, a box plot can also be used to visually describe the spread of a sample and can help identify possible outlying values—that is, values that seem inconsistent with the rest of the points in the sample. In the context of box plots, outlying values are defined as follows:

DEFINITION 2.11 An **outlying value** is a value x such that either

(1) $x >$ upper quartile + 1.5 × (upper quartile – lower quartile) or

(2) $x <$ lower quartile – 1.5 × (upper quartile – lower quartile)

DEFINITION 2.12 An **extreme outlying value** is a value x such that either

(1) $x >$ upper quartile + 3.0 × (upper quartile – lower quartile) or

(2) $x <$ lower quartile – 3.0 × (upper quartile – lower quartile)

Question: How to construct a box plot?

To construct the box plot, the top of the box corresponds to the upper quartile (Q3), whereas the bottom of the box corresponds to the lower quartile (Q1). A horizontal line is also drawn at the median value (Q2). The box plot is then completed by:

(1) Drawing a vertical bar from the upper quartile (Q3) to the largest non-outlying value (Max) in the random sample.

(2) Drawing a vertical bar from the lower quartile (Q1) to the smallest non-outlying value (Min) in the random sample.

(3) Individually identifying the outlying and extreme outlying values in the random sample by zeroes (0) and asterisks (*), respectively.

Example

Consider the data set in Table 2.9, which represents the birthweights from 100 consecutive deliveries at a Boston hospital. Using the statistical package MINITAB, answer the following:

1) Calculate the mean, median and mode?
2) Calculate the range, variance, standard deviation and coefficient of variation?
3) Calculate the lower quartile (Q1), the median (Q2) and the upper quartile (Q3)?
4) Display these data by using the box plot?
5) What is the approximate shape of the distribution of birthweights from box plot?
6) Using the box plot, comment on the spread of the sample in Table 2.9 and the presence of outlying values?

Solution

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

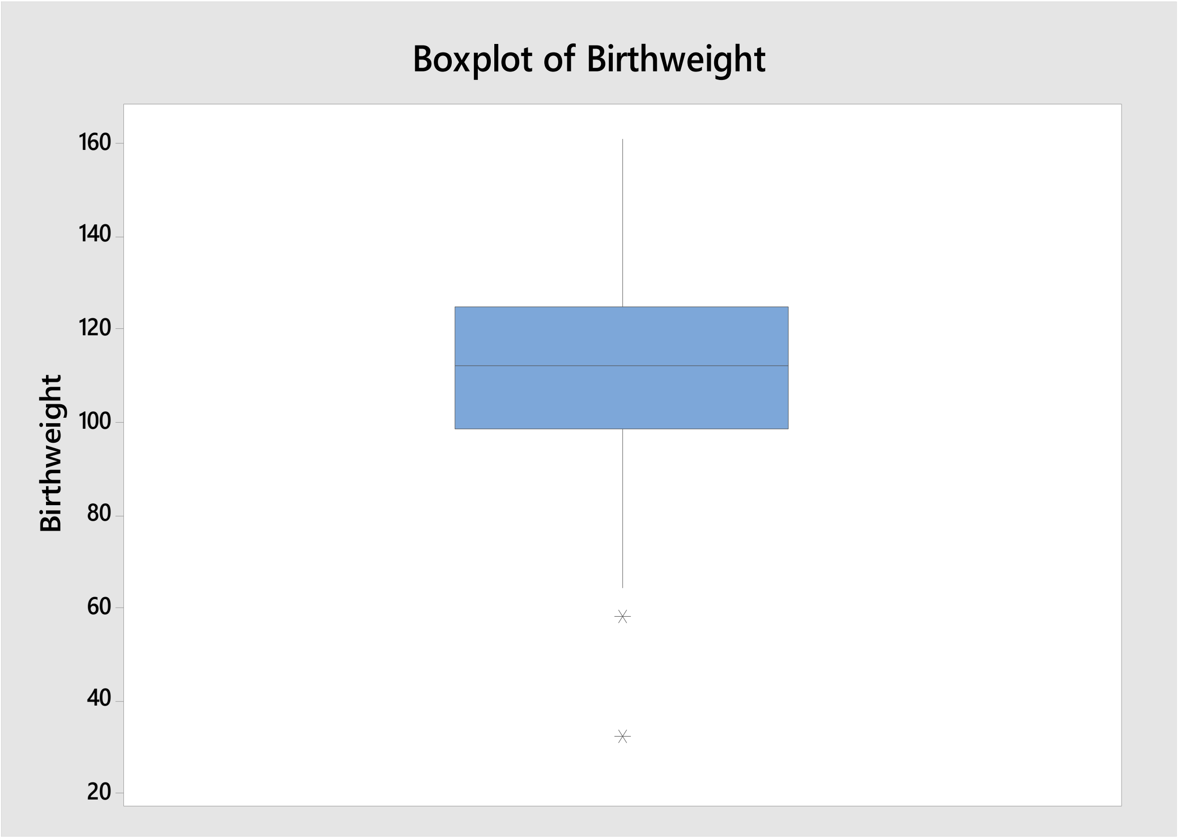
58	118	92	108	132	32	140	138	96	161
120	86	115	118	95	83	112	128	127	124
123	134	94	67	124	155	105	100	112	141
104	132	98	146	132	93	85	94	116	113
121	68	107	122	126	88	89	108	115	85
111	121	124	104	125	102	122	137	110	101
91	122	138	99	115	104	98	89	119	109
104	115	138	105	144	87	88	103	108	109
128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

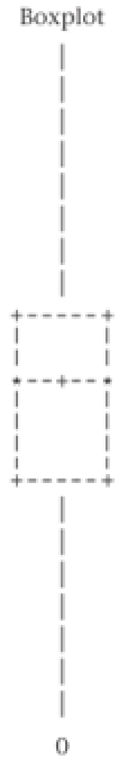
1) Answer: mean = 111.26, median = 112 and mode = 115.

2) Answer: range = 129, variance = 438.96 , standard deviation = 20.95 and coefficient of variation = 18.83.

3) Answer: lower quartile (Q1) = 98.50
median (Q2) = 112
upper quartile (Q3) = 124.50

4) Answer: Box plot





5) Answer: Approximate Shape

Because the lower quartile (Q1) is farther from the median (Q2) than the upper quartile (Q3), that is:

(Q2 – Q1 = 112 - 98.50 = 13.50 > Q3 – Q2 = 124.50 – 112 = 12.50)

then the distribution is slightly negatively skewed. This pattern is true of many birthweight distributions.

6) Answer: Spread of the sample in Table 2.9 and the Presence of Outlying Values

It can be shown from Definition 2.6 that the upper and lower quartiles are 124.5 and 98.5 oz, respectively. Hence, an outlying value x must satisfy the following relations:

$$x > 124.5 + 1.5 \times (124.5 - 98.5) = 124.5 + 39.0 = 163.5$$

or

$$x < 98.5 - 1.5 \times (124.5 - 98.5) = 98.5 - 39.0 = 59.5$$

Similarly, an extreme outlying value x must satisfy the following relations:

$$x > 124.5 + 3.0 \times (124.5 - 98.5) = 124.5 + 78.0 = 202.5$$

or

$$x < 98.5 - 3.0 \times (124.5 - 98.5) = 98.5 - 78.0 = 20.5$$

Conclusion

Thus, the values 32 and 58 oz are outlying values but not extreme outlying values. These values are identified by *'s on the box plot.

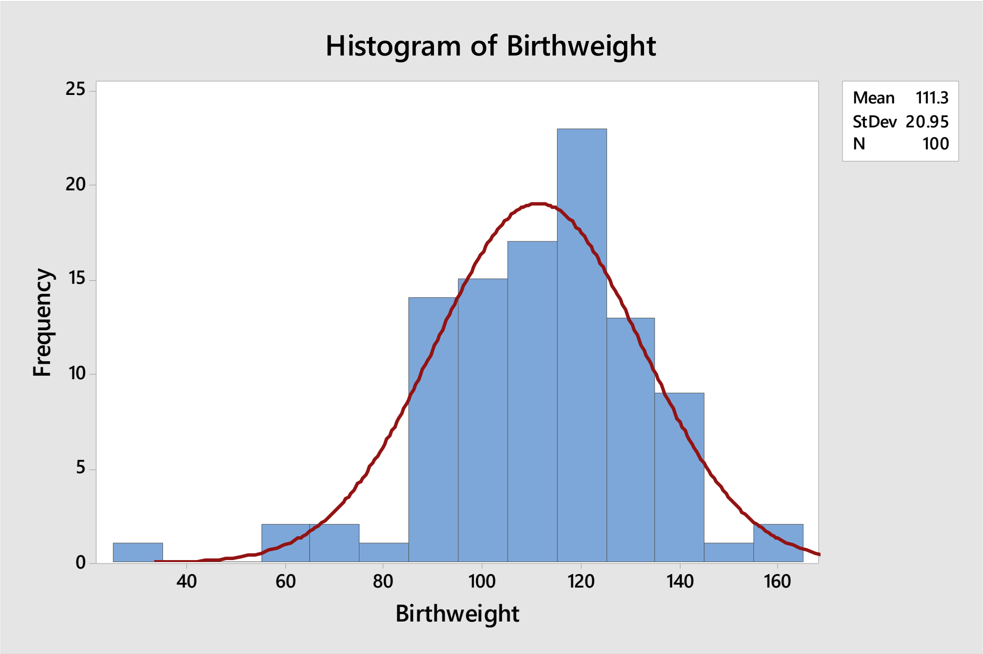
Histogram

It is a graphical display of data using bars of different heights. It is similar to a Bar Chart, but a histogram groups numbers into classes or intervals. The height of each bar shows how many fall into each class.

Example

Consider the data set in Table 2.9, which represents the birthweights from 100 consecutive deliveries at a Boston hospital. Using the statistical package MINITAB, draw the Histogram?

Solution



Problems: 2.1-2.5, 2.7-2.10, 2.12-2.18. (Use MINITAB)

The End