

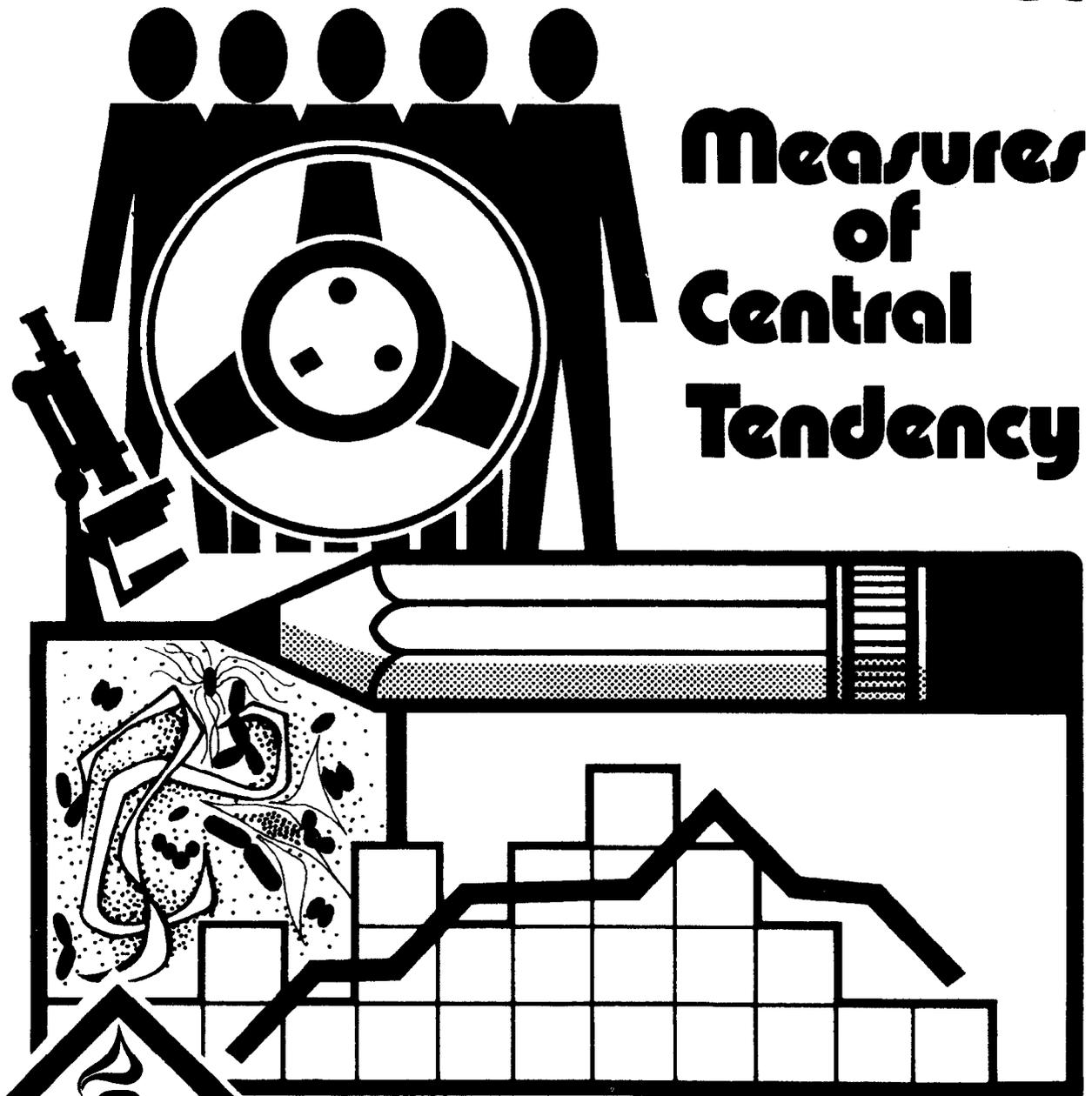


MANUAL

2

SELF-STUDY COURSE 3030-G

Principles of Epidemiology



Measures of Central Tendency

SELF-STUDY

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES

PUBLIC HEALTH SERVICE

Centers for Disease Control

Training and Laboratory Program Office

Division of Training

Atlanta, Georgia 30333

10/88:4R

MEASURES OF CENTRAL TENDENCY

Introduction	1
Individual Data: Calculation of the Mean, Median, and Mode	2
Grouped Data: Calculation of the Mean, Median, and Mode	5
Simple Frequency Distributions	5
Frequency Distributions Using Class Intervals	8
Exercises in the Calculation of the Mean, Median, and Mode	12
Appendix A: Glossary of Terms	19
Appendix B: Rounding Off Decimals	23
Appendix C: Identification of the Mid-Point of an Interval	24

PRINCIPLES OF EPIDEMIOLOGY

Self-Study Course 3030-G

MEASURES OF CENTRAL TENDENCY

INTRODUCTION

Often in epidemiologic investigations the investigator must interpret and compare information of a similar nature that has been obtained from a large number of sources. Examples would be the ages of persons who have an illness and the incubation periods of these persons' illness. Frequently, too, the investigator must compare the characteristics of two or more groups, such as the ages of people having and not having a particular illness. These interpretations and comparisons are greatly facilitated if the investigator can use, for each group of data, a single value that is in some way representative of the individual values in the group. This single, representative, value is a measure of the central tendency of the values in the group.

In epidemiologic studies measures of central tendency are commonly used to identify and describe the distribution of a group of cases according to:

- variables of place and person
- their incubation periods.
- their times of onset of illness.

These distributions provide the factual basis for hypotheses of the agent and its source and mode of transmission.

An example is comparing the ages of people who become ill subsequent to eating a particular meal to the ages of people who also ate the meal but did not become ill. It is evident the representative values for the two sets of data make interpretations and comparisons much easier:

Individual values:

- ill persons: 8, 12, 17, 7, 9, 11, 6, 3, and 13 years.
- well persons: 19, 33, 7, 26, 21, 36, 33, and 24 years.

Representative values (in this case, the arithmetic mean of the respective groups):

- ill persons: 10 years.
- well persons: 25 years.

The three most commonly-used measures of central tendency in the practice of epidemiology are the arithmetic mean, the median, and the mode (see also the Glossary, Appendix A). If there are only a few values on a set, then these measures are calculated from individual data; if there are many values in a set, then these measures are usually calculated from grouped data.

INDIVIDUAL DATA: Calculation of the Mean, Median, and Mode

The Mean

The mean is the average value of the values in a set. It is calculated by dividing the sum of the individual values in a set by the number of values in the set. Mathematically, this process is expressed as follows:

$$\bar{X} = \frac{\sum x_i}{n}$$

\bar{X} = The arithmetic mean

Σ = The sum of

x_i = Each individual value in a set

n = The number of individual values in the set

EXAMPLE: Calculation of the mean of a set of 5 values: 7, 9, 11, 13, 15

$$\bar{X} = \frac{7 + 9 + 11 + 13 + 15}{5} = \frac{55}{5} = \underline{11}$$

The arithmetic mean, because it is required for more elaborate statistical analyses, is the most commonly used measure of central tendency. A limitation of the mean is that it, of the three measures discussed here, is the measure most affected by the presence in the set of a few extreme values, large or small.

The Median

The median in a set of data is that value or point in a series which divides the ranked values into two equal-sized groups, one consisting of values equal to or smaller than the median, the other consisting of values equal to or larger than the median. The median is a more representative measure of central tendency than is the mean in those sets of data that are skewed in one direction or another (have one or more extreme values).

To obtain the median value of ungrouped data it is necessary to:

1. Rank-order, or sequence, the values in either ascending or descending order.
2. Identify the mid-point of the sequence:
 - a. If there is an odd number of values, identify the middle number;
 - b. If there is an even number of values, identify the mid-point between the two numbers in the middle of the sequence.

The general formula for identifying the mid-point, or median case, of a sequence of values is:

$$\text{median case} = \frac{\text{total number of values in the sequence} + 1}{2}$$

3. Take the value of the mid-point, or the median case, as being the median value in the set.

EXAMPLE: Median of a set containing an odd number of values:

1. Values ranked in ascending order: 7, 9, 11, 13, 15.
2. Mid-point of the sequence = $(5 \text{ values} + 1) \div 2 = 6 \div 2 = 3$
Therefore the median is the value of the 3rd value in the sequence.
3. The 3rd value is 11, so the median is 11.

EXAMPLE: Median of a set containing an even number of values:

(a) Situation where the middle numbers are different

1. Values ranked in ascending order: 7, 9, 11, 13, 15, 17
2. The median case = $(6 + 1) \div 2 = 7 \div 2 = 3.5$
Therefore the median is that value which lies midway between the 3rd and the 4th value in the sequence.
3. The 3rd and 4th values are, respectively, 11 and 13.
The median = $(11 + 13) \div 2 = \underline{24} \div 2 = \underline{12}$

(b) Situation where the middle numbers are the same

1. Values ranked in ascending order: 7, 9, 11, 11, 13, 15
2. In this example the median case is also the 3.5th case in the rank-ordered set: $(6 + 1) \div 2 = 7 \div 2 = 3.5$
3. Since the third and the fourth values of the sequence are both 11 the median is also 11: $(11 + 11) \div 2 = 22 \div 2 = \underline{11}$

The Mode

The mode is that value in a set of data which occurs most frequently. It is identified by counting the number of times each value occurs in the set and selecting that value which occurs most often. It sometimes happens that a set of data contains more than one mode. While such an event may complicate the interpretation of the data, it may also have epidemiologic significance. Because modes generally cannot be used in more elaborate statistical calculations, they are used less frequently than means.

EXAMPLE: Mode of a set of 7 observations:

3, 4, 6, 6, 6, 9, 12

The mode is 6 since it is the one which occurs most frequently (3 times).

EXAMPLE: A set of 12 observations containing more than one mode:

1, 3, 4, 4, 4, 5, 7, 9, 9, 9, 14, 15

There are two numbers that occur with equal frequency, so this distribution has two modes--4 and 9.

As another example of the use and method of calculation of these three measures, consider the following set of data which represent the ages in years of persons involved in a small outbreak: 1, 7, 3, 2, 3, 39, 4, 8, 3, 24. Calculation of the mean, median, and mode of the ages of the cases involved is as follows:

1. Mean age = (Sum of ages of all cases) ÷ (Total number of cases)
or, mathematically:

$$\bar{X} = \frac{\sum x_i}{n} \quad \sum x_i = 1 + 7 + 3 + 2 + 3 + 39 + 4 + 8 + 3 + 24 = 94$$

$$n = 10$$

Substituting these numbers for the symbols, we have:

$$\bar{X} = \frac{\sum x_i}{n} = \frac{94}{10} = \underline{9.4 \text{ years}}$$

NOTE: See Appendix B for the method used for rounding off numbers.

2. Median age:

(a) Rank-ordered values: 1, 2, 3, 3, 3, 4, 7, 8, 24, 39

(b) Mid-point: $(10 + 1) \div 2 = 11 \div 2 = 5.5$
Therefore, the median lies midway between the fifth and sixth values in the sequence.

(c) The fifth and sixth values in the sequence are, respectively, 3 and 4. Therefore the median = $(3 + 4) \div 2 = 7 \div 2 = \underline{3.5}$

3. Modal age: By counting the number of times each value occurs in the set we can identify 3 as being the modal value:

Age:	1	2	3	4	7	8	24	39
Frequency:	/	/	///	/	/	/	/	/

So in this set of data we have the following measures of central tendency:

Mean age of cases = 9.4 years
Median age of cases = 3.5 years
Modal age of cases = 3.0 years

Which of the above provides the best (i.e., most representative) measure of central tendency for this set of data? Probably the median since there is a skewing of the distribution to the older ages: among the cases are two (24 & 39) that could be considered to have an extreme value with respect to the other observations. The effect of these two extreme values on the mean can be easily seen.

GROUPED DATA: Calculation of the Mean, Median, and Mode

Simple Frequency Distribution

The discussion and calculations up to this point have been on the identification of the mean, median, and mode for sets of individual observations. Means, medians, and modes also can be obtained from data tabulated into a simple frequency distribution. This is often done to facilitate the calculation for these measures. A "simple" frequency distribution is one based on single units of whatever measurement units are being used. In table 1 the measurement unit is hours, and the cases are tabulated according to the specific number of hours that lapsed between the time the case was exposed to the agent and the time at which the case became ill.

Table 1

Incubation Periods of 41 Cases of Disease "A" Among Persons in a Common Source Outbreak, Sample City, June 2, 1976

Col. (1) Incubation Period in Hours (class interval)	Col. (2) Number of Cases (f_i)
8- 8.9	2
9- 9.9	4
10-10.9	9
11-11.9	14
12-12.9	7
13-13.9	2
14-14.9	1
15-15.9	0
16-16.9	2
TOTAL	41

Source: Epidemic Report, Sample City Health Department.

The mean of data tabulated as in Table 1 (i.e., a simple frequency distribution) is calculated using this formula:

$$\bar{X} = \frac{\sum f_i x_i}{n}$$

The symbols are defined as:

\bar{X} = The arithmetic mean

Σ = the sum of

f_i = the frequency of occurrence of an event (e.g., the numbers in column 2 of Table 1)

x_i = the measurement units or the mid-points of the intervals (e.g., the numbers in column 3 of Table 2)

n = the number of individual values in the set

In order to calculate the mean incubation period of the cases of disease "A" shown in Tables 1 and 2, it is necessary to add a third and fourth columns to the table. The third column will contain the mid-point of each incubation period shown in column 1, and the fourth column will contain the individual products of the multiplication of each mid-point in column 3 by the figure on the same line in column 2. Column 4 is then totaled. For a discussion of how mid-points are identified see Appendix C. Table 2 shows the results.

Table 2

Incubation Periods of 41 Cases of Disease "A" Among Persons in a Common Source Outbreak, Sample City, June 2, 1976

Col. (1) Incubation Period in Hours (class interval)	Col. (2) Number of Cases (f_i)	Col. (3) Mid-point of Interval Shown (x_i)	Col. (4) Products of Columns 1 and 2 ($f_i x_i$)
8- 8.9	2	8.5	17.0
9- 9.9	4	9.5	38.0
10-10.9	9	10.5	94.5
11-11.9	14	11.5	161.0
12-12.9	7	12.5	87.5
13-13.9	2	13.5	27.0
14-14.9	1	14.5	14.5
15-15.9	0	15.5	0.0
16-16.9	2	16.5	33.0
TOTAL	41	-----	472.5

Now the data can be put into the formula and the mean calculated:

from the table we have: $\sum f_i x_i = 472.5$

$$n = 41$$

$$\bar{X} = \frac{\sum f_i x_i}{n} \quad \bar{X} = \frac{472.5}{41} = \underline{11.5 \text{ hours}}$$

Now the median incubation period. As in individual data, the first step is to identify the middle case. Middle case = $\frac{\text{Total cases} + 1}{2}$. Since there are 41 cases the 21st case is the middle case $\frac{41 + 1}{2} = \frac{42}{2} = 21$. Now we need to find the class interval (incubation period) in which the 21st sequential case occurs. To do this we simply accumulate the number of cases (in column 2) starting with the shortest incubation period, until the interval is reached that contains the middle case as in Table 3.

Table 3

Incubation Periods of 41 Cases of Disease "A" Among Persons in a Common Source Outbreak, Sample City, June 2, 1976

(Col. 1) Incubation Period in Hours (class interval)	Col. (2) Number of Cases (f_i)	Col. (3) Accumulation of Column 2 ($\sum f_i$)
8- 8.9	2	2
9- 9.9	4	6
10-10.9	9	15
11-11.9	14	29 (21st observation falls in this interval)
12-12.9	7	36
13-13.9	2	38
14-14.9	1	39
15-15.9	0	39
16-16.9	2	41
TOTAL	41	

The 21st case is included in the range of 15 to 29 cases, which occurred in the 11-hour interval. Therefore, the median incubation period is 11 hours.

The mode, as with individual data, is simply that incubation period in which the most cases occurred. Inspection of column 2 in Table 3 reveals that the largest number of cases in any single hourly incubation period is 14, and it occurred in the period of 11 - 11.9 hours. Therefore, the modal incubation period is 11 hours.

Frequency Distribution Using Class Intervals

Suppose that the information shown in Table 1 had been organized differently, so that the cases were tabulated by 2-hour rather than 1-hour incubation period intervals (or "class intervals"; see Glossary), as is shown in Table 4.

Table 4

Incubation Periods of 41 Cases of Disease "A" Among Persons in a Common Source Outbreak, Sample City, June 2, 1976

Col. (1) Incubation Period in Hours (class interval)	Col. (2) Number of Cases (f_i)
8- 9.9	6
10-11.9	23
12-13.9	9
14-15.9	1
16-17.9	2
TOTAL	41

Source: Epidemic Report, Sample City Health Department

The mean, median, and mode can still be obtained, with no modifications to our previous methods.

The mean

The formula to use is the same as previously:

$$\bar{X} = \frac{\sum f_i x_i}{n}$$

\bar{X} = the arithmetic mean

Σ = the sum of

f_i = the frequency of events in respective intervals

x_i = mid-point of the class interval (see Appendix C)

n = total number of events

It is necessary now (as in the previous example) to expand Table 4 into a new table (Table 5) by adding two columns, one (column 3) showing the mid-point of each class interval, and one (column 4) showing the product of the mid-point (column 3) and the respective frequency of occurrence (column 2).

Table 5

Incubation Periods of 41 Cases of Disease "A" Among Persons in a Common Source Outbreak, Sample City, June 2, 1976

Col. (1) Incubation Period in Hours (class intervals)	Col. (2) Number of Cases (f_i)	Col. (3) Mid-point of Class Interval (x_i)	Col. (4) Product of Columns 2 and 3 ($f_i x_i$)
8- 9.9	6	9	54
10-11.9	23	11	253
12-13.9	9	13	117
14-15.9	1	15	15
16-17.9	2	17	34
TOTAL	41	-----	473

So, substituting numbers for symbols in the formula, we get:

$$\bar{X} = \frac{\sum f_i x_i}{n} = \frac{473}{41} = \underline{11.5 \text{ hours}}$$

The calculation of a mean from values tabulated in a frequency distribution using class intervals is based on an important assumption: that the values in each class interval are at, or evenly distributed about, the mid-point of the respective class interval. An important consequence of this is that class intervals must be selected which are not too large with respect to the entire range of values and the distribution of the values within the range.

The Median

In calculating the median of a set of data grouped in class intervals, the assumption is made that the individual values within each class interval are equally distributed throughout that interval. In Table 5, for example, the assumption is that the 6 cases which had an incubation period of 8-9.9 hours occurred uniformly at 20 minute intervals throughout the two-hour period (2 hours \div 6 cases = 120 minutes \div 6 cases = 20 minutes each).

To calculate the median of the data in Table 5, proceed as follows:

1. Identify the middle case. This has already been calculated to be the 21st case:
 $(41 + 1) \div 2 = 42 \div 2 = \underline{21}$
2. Identify the class interval that contains the middle case by accumulatively totaling the number of cases in each class interval until the class interval is reached that contains the middle case. The middle case, number 21, falls between 6 and 29, the accumulated totals of the first and second intervals, respectively. Therefore, the 21st case occurs in the second (i.e., the 10-11.9 hour) interval.
3. Identify the total number of the first 21 cases that fall in this interval:
number = 21 - (the number of cases in preceding intervals)
= 21 - 6 = 15. Therefore, 15 of the 21 cases "belong" to the 10-11.9 hour interval.
4. Calculate the time required for the first 15 cases to occur in the second (10-11.9 hours) interval (again, assuming that all of the 23 cases in that interval occurred uniformly during the interval):

Time required = (15 cases x 2 hours) \div 23 cases in the interval
= 30 \div 23 = 1.30 = 1.3 hours
5. Add this portion to the beginning of the interval to get the median:
10 hours + 1.3 hours = 11.3 hours

Therefore, the median incubation period of the cases in this outbreak was 11.3 hours.

Suppose though, that the data set contained an even number of values. How would you calculate the median? If, in Table 5, the incubation period interval of 14-15.9 hours had no cases, the "n" would be 40 and $\sum f_i x_i$ would be 458 (473 - 15 = 458). Following the same 5 steps as before, the median would be calculated like this:

1. Median case = $(n + 1) \div 2 = (40 + 1) \div 2 = 41 \div 2 = \underline{20.5}$
2. Measurement interval containing the median case: 10-11.9 hours (determined the same as before)
3. Portion of the 20.5 cases falling in the 10-11.9 hour interval = 20.5 - (number of cases in the preceding intervals) = 20.5 - 6 = 14.5 cases

4. Portion of the 2-hour interval equivalent to 14.5 cases = $(14.5 \times 2 \text{ hours}) \div 23 \text{ cases in that interval} = 29 \div 23 = 1.26 = \underline{1.3 \text{ hours}}$
5. Add the 1.3 hours to the start of the interval, 10-11.9 hours, to get the median incubation period:
 $10 + 1.3 = \underline{11.3 \text{ hours}}$

The Mode

The mode remains that class interval having the greatest frequency of events. In Table 5, the mode is 10-11.9 hours. Clearly the mode in grouped data is an imprecise figure, and this imprecision increases as the size of the class interval increases. Somewhat more precision can be obtained by following the common practice of taking the mode to be the mid-point of the interval having the greatest frequency of events. In Table 5, this procedure would provide a mode of 11 hours.

NOTE:

ADDITIONAL EXAMPLES OF HOW TO CALCULATE THE MEAN, MEDIAN, AND MODE FROM INDIVIDUAL AND GROUPED DATA ARE PROVIDED ON PAGES 12 THROUGH 17. These EXAMPLES ARE BASED ON DATA REGARDING:

- DURATION OF ILLNESS
- INCUBATION PERIODS
- AGE GROUPS OF CASES

THE FIRST THREE EXAMPLES SHOW THE METHOD OF CALCULATION OF THE VARIOUS MEASURES; ADDITIONAL EXAMPLES SHOW ONLY THE ANSWER TO THE PROBLEMS POSED.

IT WOULD BE HIGHLY DESIRABLE FOR YOU TO WORK THROUGH ALL OF THESE EXAMPLES.

EXERCISES IN THE CALCULATION OF SELECTED MEASURES OF CENTRAL TENDENCY

Part A: Exercises in which the Method of Calculation is shown

Following are three sets of data. The first set consists of individual (ungrouped) data; the second consists of a simple frequency distribution (grouped) data; and the third consists of a frequency distribution using class intervals (grouped) data. For each set of data the mean, median, and mode are to be calculated.

EXERCISE 1

Given: Duration of illness, in days, among cases having disease "A":
9, 7, 11, 9, 8, 4, 6, 12, 6, 8, 8, 5

Required: Calculate the mean, median, and mode duration of illness

Solution:

a. Mean = $\frac{\sum x_i}{n} = \frac{9 + 7 + 11 + 9 + 8 + 4 + 6 + 12 + 6 + 8 + 8 + 5}{12}$
 $= \frac{93}{12} = 7.75 = \underline{7.8 \text{ days}}$

b. Median:

1. middle case = $\frac{n + 1}{2} = \frac{12 + 1}{2} = \frac{13}{2} = 6.5$

Which means that the median is that value mid-way between the 6th and the 7th cases.

2. The durations of illness, in ascending order, and with the 6th and the 7th cases shown, are:

6th case } 7th case
 4, 5, 6, 6, 7, 8, 8, 8, 9, 9, 11, 12

3. The middle number, falling between two 8's, is 8 days.

c. Mode:

1. Tabulation of cases according to their durations:

Duration of illness, in days	=	4	5	6	7	8	9	11	12
Number of cases of each duration	=	/	/	//	/	///	//	/	/

2. The greatest number of cases of any single duration is three, and those occurred on the eighth day; therefore, the mode is 8 days.

EXERCISE 2

Given: The frequency distribution of the incubation periods, in single days, of 84 cases of a disease:

Col. (1) Incubation Period in Days (x_i)	Col. (2) Number of Cases (f_i)
16-16.9	1
19-19.9	1
23-23.9	2
24-24.9	4
25-25.9	3
26-26.9	8
27-27.9	17
28-28.9	15
29-29.9	16
30-30.9	9
31-31.9	4
32-32.9	2
36-36.9	1
41-41.9	1
TOTAL	84

Required: Calculate the mean, median, and mode incubation period:

Solution: Expanding the preceding table to obtain the additional data needed to calculate the mean and median yields:

Col. (1) Incubation Period in Days	Col. (2) Number of Cases (f_i)	Col. (3) Mid-point of Interval (x_i)	Col. (4) Product of Col. (2) and Col. (3) ($f_i x_i$)	Col. (5) Cumulative Number of Cases in Col. (2) (Σf_i)
16-16.9	1	16.5	16.5	1
19-19.9	1	19.5	19.5	2
23-23.9	2	23.5	47.0	4
24-24.9	4	24.5	98.0	8
25-25.9	3	25.5	76.5	11
26-26.9	8	26.5	212.0	19
27-27.9	17	27.5	467.5	36
28-28.9	15	28.5	427.5	51
29-29.9	16	29.5	472.0	67
30-30.9	9	30.5	274.5	76
31-31.9	4	31.5	126.0	80
32-32.9	2	32.5	65.0	82
36-36.9	1	36.5	36.5	83
41-41.9	1	41.5	41.5	84
TOTAL	84	-----	2380	-----

a. Mean: $\bar{X} = \frac{\Sigma f_i x_i}{n} = \frac{2380}{84} = 28.33$

mean incubation period = 28.3 days

b. Median: The middle case = $\frac{n+1}{2} = \frac{84+1}{2} = \frac{85}{2} = 42.5$

Reviewing column 5 it can be seen that the 42nd and 43rd cases both occur on the 28th day. Therefore, the median incubation period = 28 days.

c. Mode: The highest frequency of cases (17) occurred on the 27th day, making it the mode.

EXERCISE 3

Given: Age group distribution of cases who were involved in an outbreak.

Age Group in Years	Number of Cases
0- 0.9	0
1- 4.9	4
5- 9.9	9
10-19.9	14
20-29.9	13
30-59.9	3
60-99.9	0
TOTAL	43

Required: Calculate the mean, median, and mode age.

Solution: Expanding the preceding table to obtain the additional data needed to calculate the mean and the median yields:

Col. (1) Age Group in Years (class interval)	Col. (2) Mid-point of Class Interval (x_i)	Col. (3) Number of Cases (f_i)	Col. (4) Product of Col. (2) and Col. (3) ($f_i x_i$)	Col. (5) Accumulation of Col. (3) (Σf_i)
0- 0.9	0.5	0	0.0	0
1- 4.9	3.0	4	12.0	4
5- 9.9	7.5	9	67.5	13
10-19.9	15.0	14	210.0	27
20-29.9	25.0	13	325.0	40
30-59.9	45.0	3	135.0	43
60-99.9	80.0	0	0.0	43
TOTAL	-----	43	749.5	----

a. Mean: $\bar{X} = \frac{\Sigma f_i x_i}{n} = \frac{749.5}{43} = 17.43 = \underline{17.4 \text{ years}}$

b. Median:

1. Middle case = $\frac{n + 1}{2} = \frac{43 + 1}{2} = \frac{44}{2} = 22\text{nd case.}$
2. 22nd case is located in 10 - 19.9 age group.
3. 9, (22 - 13 = 9), cases of the 22 fall in the 10 - 19.9 age group.
4. $\frac{9}{14} \times 10 \text{ yrs.} = 6.42 - 6.4 \text{ years.}$
5. 10 years + 6.4 years = 16.4 years.

c. Mode:

1. Interval having greatest frequency, 14 cases = 10 - 19 age group.
2. Mid-point of modal interval = 15 years.

Part B: Exercises in which only the answers are provided

The following exercises are intended to provide you with an opportunity to further practice the identification of the mean, median, and mode of a set of data. Several sets of data are provided. You should identify the mean, median, and mode of each set and compare your answers with those provided on the next page.

Exercise 1. 1, 7, 9, 14, 15, 21

Exercise 2. 6, 8, 12, 12, 14, 18, 26

Exercise 3. 20, 20, 20, 21, 27, 32, 33, 40

Exercise 4

Incubation Period in Hours (x_i)	Number of Cases (f_i)
8- 8.9	2
9- 9.9	1
10-10.9	6
11-11.9	18
12-12.9	11
13-13.9	4
14-14.9	1
15-15.9	1

Exercise 5

Incubation Period in Hours (x_i)	Number of Cases (f_i)
1-1.9	0
2-2.9	6
3-3.9	12
4-4.9	8
5-5.9	2
6-6.9	1
7-7.9	0
8-8.9	0

Exercise 6

Incubation Period in Hours (class interval)	Number of Cases (f_i)
6- 7.9	3
8- 9.9	7
10-11.9	17
12-13.9	16
14-15.9	12
16-17.9	6

Exercise 7

Age of Cases in Years (class interval)	Number of Cases (f_i)
0- 4.9	5
5-19.9	0
20-39.9	2
40-64.9	7
65-99.9	21

ANSWERS TO PRECEDING EXERCISES

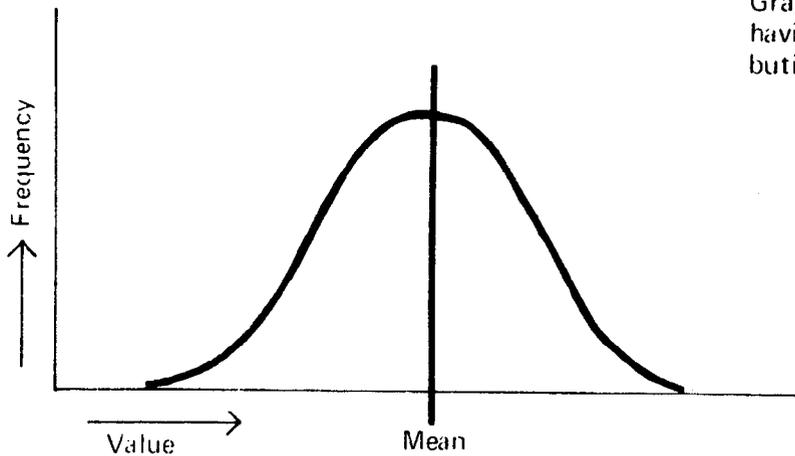
Exercise Number	Answer		
	Mean	Median	Mode
1	11.2	11.5	none
2	13.7	12	12
3	26.6	24	20
4	11.8 hours	11 hours, or 11–11.9 hours	11 hours, or 11–11.9 hours
5	3.8 hours	3 hours, or 3–3.9 hours	3 hours, or 3–3.9 hours
6	12.5 hours	a. interval = 12–13.9 hours b. exact time = 12.5 hours	a. modal interval = 10–11.9 hours b. mid-point of the modal interval = 11 hours
7	62.1 years	a. interval = 65–99.9 years b. exact age = 71.7 years	a. modal interval = 65–99.9 years b. mid-point of the modal interval = 82.5 years

APPENDIX A
GLOSSARY OF TERMS

GLOSSARY OF TERMS

- Biased events** Events whose occurrence is either encouraged or discouraged by one or more circumstances; non-random.
- Class interval** In a table prepared for the purpose of obtaining a frequency distribution of the values in a set, a class interval is a sub-range of values within a larger range. The length of the intervals that make up the range do not have to be the same size, but the limits of adjacent intervals must be mutually exclusive:
- correct: 0-4, 5-9, 10-14, 15-19,
20-49, 50-79, 80+
- incorrect: 0-4, 4-9, 9-14, 14-19, etc.
- incorrect: 0-5, 5-10, 10-15, 15-20,
etc.
- Grouped data** Data that have been tabulated into a frequency distribution.
- Measure of central tendency** A value that attempts to represent a group of individual values in a simple and concise manner. Some common measures of central tendency are the mean, the median, and the mode.
- Mean The sum of the individual values in a set divided by the number of values in that set.
- Median The middle value in a rank-ordered set containing an odd number of values; the mid-point in a similar set containing an even number of values.
- Mode The value that occurs most frequently in a set.

Normal distribution A frequency distribution in which there is one peak, centered on the mean, around which the individual values are symmetrically distributed.



Graph of a set of data having a normal distribution.

Random events Events which occur by chance alone. They usually represent the population from which they were drawn, which may or may not be normally distributed.

Range of values In a set of values, the range is the difference between the highest and lowest values. A description of the range should also specify the highest and lowest values in the set.

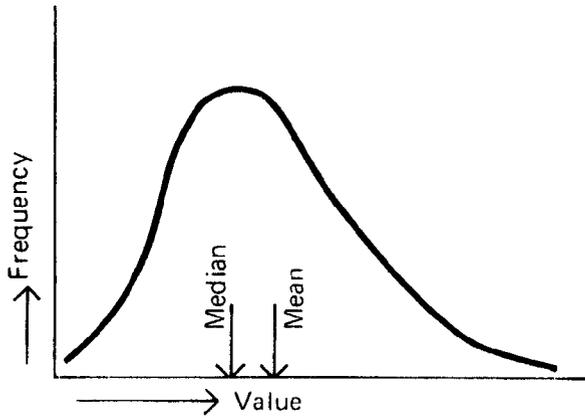
Ranked (rank-ordered) values A set in which the individual values have been arranged in sequence from the lowest value in the set to the highest value in the set. The following set is properly ranked: 6, 7, 9, 11, 17, 23, 24, 31. The ranked values may be in either ascending or descending order.

Set A group of data in which the individual values are related in some way (e.g., ages of individuals in a population, ages of individuals in an outbreak, the incubation periods of these cases).

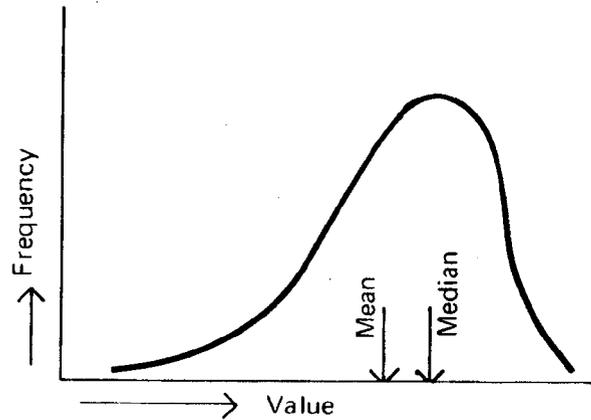
Skewed distribution A frequency distribution which is asymmetrical. Such distributions are described as skewed to the right or to the left.

GRAPHS OF SETS OF DATA WHICH ARE SKEWED:

a. Skewed to the right



b. Skewed to the left



APPENDIX B

ROUNDING OFF DECIMALS

For the purposes of this course answers are calculated to 2 digits to the right of the decimal point and rounded off to one digit. In the examples that follow, the first number is the answer to a calculation carried to 2 digits past the decimal, and the second number is the correctly rounded answer.

	Initial Calculation		Calculation Correctly Rounded Off
1.	86.74	=	86.7
2.	171.66	=	171.7
3.	413.97	=	414.0
4.	2.17	=	2.2
5.	276.94	=	276.9
6.	76.02	=	76.0
7.	337.55	=	337.6

The rule followed in these examples is that if the second digit to the right of the decimal point is 4 or less it is simply dropped, and the first digit to the right of the decimal remains unchanged (as in examples 1, 5, and 6). If the second digit to the right of the decimal point is 5 or greater, then the first digit to the right of the decimal is increased by one (examples 2, 3, 4, and 7).

APPENDIX C

IDENTIFICATION OF THE MID-POINT OF AN INTERVAL

The arrangement of the individual values in a set of data into a frequency distribution clearly has some uses. An important decision to be made before arranging individual values into a frequency distribution is the selection of the size of the intervals by which the data are to be tabulated. The size of the intervals to be used depends upon the questions you are trying to answer with the data, and the number, range, distribution and precision of the values in the set.

For example, suppose you wish to obtain the frequency distribution of the ages of children having disease "A". If the number of children involved was very small, say less than 10, you might not make a frequency distribution at all. You might simply list the ages in order of increasing size. If a larger number of children were involved, though, say between 10 and 50, you would need to prepare a frequency distribution. Commonly, exact ages of cases are not obtained or necessary; instead the age at the last birthday is obtained. In this instance, the frequency distribution of the children could be according to either individual year of age, or age group. Considering the former first, the frequency distribution probably would be as follows in Table 1.

Table 1

Age (in years at last birthday) (x_i)	Number of Cases (f_i)
0	0
1	0
2	6
3	12
4	17
,	,
,	,
,	,
etc.	---

In this kind of distribution (simple frequency distribution) where the span of the variable is measured in single units of a variable (in this case, single years) a mid-point commonly is not calculated. Instead the unit itself is multiplied by its respective frequency (i.e., the number of cases having that age).

However, if exact ages were obtained, and the additional precision was necessary, then it would be more appropriate to prepare the frequency distribution using age intervals such as in Table 2 below. In instances where the data are tabulated by intervals, as they are here, then the mid-point (x_i) of each interval must be identified. In Table 2, the mid-point of the interval 0.00 - 0.99 is obtained by adding the upper limit (0.99) to the lower limit (0.00) and dividing by 2 as follows:

$$\text{Mid-point (0.00 - 0.99)} = \frac{0.00 + 0.99}{2} = \frac{0.99}{2} = \underline{0.495} = \underline{0.5}$$

In fact, this is the method of finding the mid-point of any interval discussed in this reference.

Table 2

Age, exact, in years (class interval)	Number of Cases (f_i)	Mid-point of class interval (x_i)
0.00-0.99	0	0.5
1.00-1.99	0	1.5
2.00-2.99	6	2.5
3.00-3.99	12	3.5
4.00-4.99	17	4.5
,	,	,
etc.	----	----

Other examples of class intervals are as follows in Tables 3, 4, and 5. Notice that the end of one interval and the beginning of the next interval is not and cannot be the same number. Table 6 is an example of an INCORRECT class interval.

Table 3 Correct Interval	Table 4 Correct Interval	Table 5 Correct Interval	Table 6 <u>Incorrect</u> Interval
Age (In Years)	Age (In Years)	Age (In Years)	Age (In Years)
0-1	0.00-1.99	0- 4	0- 4
2-3	2.00-3.99	5-19	4-19
4-5	4.00-5.99	20-49	19-49
6-7	6.00-7.99	50-99	49-99
,	,	,	,
,	,	,	,
etc.	etc.	etc.	etc.
TOTAL	TOTAL	TOTAL	TOTAL

In Table 3, while the ages are shown in whole years (age at last birthday) the determination of the mid-point depends upon the nature of the average desired. The mid-point for calculating the average age at last birthday would be determined as follows:

$$\text{Mid-point (0 - 1)} = \frac{0 + 1}{2} = \frac{1}{2} = \underline{0.5}$$

This is the method more commonly used in epidemiologic studies. However, if it was necessary to determine the precise average age of individuals the grouping 0-1 would actually represent a class interval of from 0.0 to 1.99 years of age as shown in Table 4. The class interval mid-points for calculating the average age would then be determined as follows:

$$\text{Mid-point (0.0 - 1.99)} = \frac{0.0 + 1.99}{2} = \frac{1.99}{2} = \underline{0.995} = \underline{1.0}$$

Referring to Table 5, having unequal class intervals, calculation of the mid-points by both methods would be as follows:

Calculation of Class Interval Mid-Point

Class Interval	For the Average Age at Last Birthday	For the Precise Average Age
(0- 4)	$\frac{4+0}{2} = \frac{4}{2} = \underline{2.0}$	$\frac{4.9+0}{2} = \frac{4.9}{2} = 2.45 = \underline{2.5}$
(5-19)	$\frac{19+5}{2} = \frac{24}{2} = \underline{12.0}$	$\frac{19.9+5}{2} = \frac{24.9}{2} = 12.45 = \underline{12.5}$
(20-49)	$\frac{49+20}{2} = \frac{69}{2} = \underline{34.5}$	$\frac{49.9+20}{2} = \frac{69.9}{2} = 34.95 = \underline{35.0}$
(50-99)	$\frac{99+50}{2} = \frac{149}{2} = \underline{74.5}$	$\frac{99.9+50}{2} = \frac{149.9}{2} = 74.95 = \underline{75.0}$

Although the discussion here has been limited to the variable of age, the method for identifying the mid-point of an interval is equally applicable to other variables, such as time or concentration of a chemical in a liquid.

UNLESS OTHERWISE SPECIFIED, EXACT MID-POINTS OF CLASS INTERVALS AS SHOWN IN TABLES 2 AND 4 ON THE PRECEEDING PAGE WILL BE USED IN THIS COURSE.