



MANUAL

4

SELF-STUDY COURSE 3030-G

# Principles of Epidemiology



## Methods for Organizing Epidemiologic Data



**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

## TABLE OF CONTENTS

### METHODS FOR ORGANIZING EPIDEMIOLOGIC DATA Tables, Graphs, and Charts

Introduction .....	1
Methods of Organizing Epidemiologic Data	
Tables .....	2
Graphs .....	8
Arithmetic Scale Line Graph .....	9
Semi-logarithmic Scale Line Graph .....	12
Histogram .....	16
Frequency Polygon .....	18
Charts .....	22
Bar Chart .....	22
Geographic Coordinate Chart .....	23
Glossary of Terms .....	31
Decision Guide to Selecting a Method of Illustrating Epidemiologic Data .....	32
Exercise on the Organization of Epidemiologic Data .....	36
Answers to the Exercise Problems .....	51

# PRINCIPLES OF EPIDEMIOLOGY

Self-Study Course 3030-G

## METHODS FOR ORGANIZING EPIDEMIOLOGIC DATA

### INTRODUCTION

The purpose of this lesson is to enable the student to identify the frequency of occurrence and the distribution of cases in a population by applying various common methods of organizing disease surveillance and investigation data according to variables of time, place, and person.

Specifically, given a set of epidemiologic data consisting of a line listing of cases and selected characteristics of each, the student will be able to select and correctly use the most appropriate method or methods of organizing that data. The criteria of successful performance are that (1) no errors have been introduced into the data during the process of its being organized; (2) true differences in the frequency and distribution of cases are identified; (3) false or misleading impressions are not conveyed by the method or techniques used; and (4) the methods used are in conformance with generally recommended practices discussed in this reference.

The methods of organizing epidemiologic data which are described in this lesson are:

1. Tables--1-, 2-, and 3- variable tables, and master tables.
2. Graphs
  - a. Arithmetic scale line graphs
  - b. Semi-logarithmic scale line graphs
  - c. Histograms--equal interval
  - d. Frequency polygons--equal interval
3. Charts
  - a. Bar charts
  - b. Maps--spot and area

Each of the above methods will be presented separately. The definition, use and features of each will be followed by the techniques used to construct that form of organization and by examples of correct usage. At the end of the lesson is a glossary (Appendix A) and a guide to the selection of an appropriate method of illustrating epidemiologic data (Appendix B), and a set of practice exercises.

Given that you have accurate, complete data, the next consideration is how to best organize it for analysis and for presentation to others. Usually information of a quantitative nature is first organized by tabulating it; and often the process stops there. This is not always, or necessarily, bad; but tabulated data does have some limitations. Foremost among these, especially when the amount of tabulated information is large, is that patterns and trends are difficult to identify and interpret. Graphs and charts are very helpful in solving this problem. With them, and assuming that the data are tabulated into meaningful groups or intervals, patterns and trends become highly visible.

Tables, graphs, and charts, have some features which are unique to each, and others which they have in common. Of primary importance among those features they have in common are the following:

1. The number of main points in any one table, graph or chart should be limited to that which can be easily understood.
2. To clearly identify the material that has been organized, a title that describes the content as to the subject, person, place, and time should always be included. Titles should be numbered.
3. To identify true differences in the frequency of occurrence and distribution of cases, the body of the data presented must be arranged in meaningful intervals of the epidemiologic variables used.
4. The source of the data should always be identified, usually as a footnote.

#### TABLES

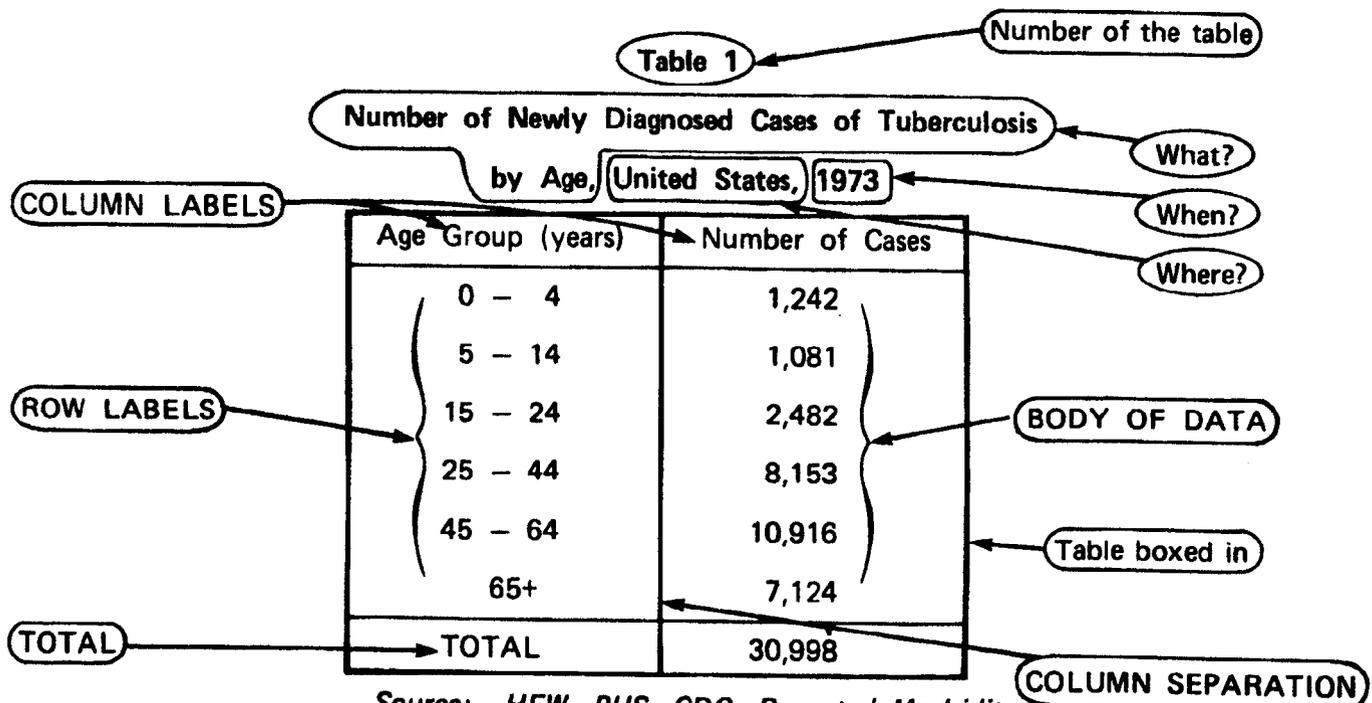
A table is a set of data that is arranged in rows and columns. The use of a table is to present the frequency with which some event occurs in different categories or subdivisions of a variable. Almost any quantitative information can be organized in tabular form, and it is the most commonly encountered method. It is from tables that graphs and charts are prepared.

Tables should be as simple as possible. Two or three small tables usually are preferred to a single large table containing many details or variables. Generally, three variables are a maximum number which can be read with ease.

Tables should be self-explanatory:

1. The title should be clear, concise, and to the point. It answers the questions of what, when and where.
2. Each row and each column should be labeled concisely and clearly. The specific units of measure for the data should be given. Columns should be separated by a vertical line.
3. Row and column totals should be shown.
4. Codes, abbreviations, or symbols should be explained in detail in a footnote.

The simplest table is a two-column (one-variable) frequency table such as Table 1. The first column lists the classes into which the data are grouped (age groups). The second column lists the frequencies for each classification (number of cases in each age group).



Source: HEW, PHS, CDC, Reported Morbidity and Mortality in the United States, 1974. Weekly Report for Year Ending December 28, 1974. Vol. 23, No. 53, page 12.

SOURCE OF INFORMATION

Table 1 can be expanded as shown in Tables 2 and 3 to include information for various sub-groups such as age, sex, or race. Table 2 is a two-variable table (age-group and race) and Table 3 is a three-variable table (age-group, race, and sex).

**Table 2**

**Number of Newly Diagnosed Cases of Tuberculosis  
by Age and Race, United States, 1973**

Age Group (years)	NUMBER OF CASES, BY RACE		
	White	Other	Total
0 - 4	674	568	1,242
5 - 14	526	555	1,081
15 - 24	1,263	1,219	2,482
25 - 44	4,017	4,136	8,153
45 - 64	6,841	4,075	10,916
65+	5,271	1,853	7,124
<b>TOTAL</b>	<b>18,592</b>	<b>12,406</b>	<b>30,998</b>

Source: (1)

**Table 3**

**Number of Newly Diagnosed Cases of Tuberculosis by  
Age, Race and Sex, United States, 1973**

Age Group (years)	NUMBER OF CASES, BY RACE AND SEX								
	White			Other			Total		
	Male	Female	Total	Male	Female	Total	Male	Female	Total
0 - 14	620	580	1,200	593	530	1,123	1,213	1,110	2,323
15 - 44	3,126	2,154	5,280	3,194	2,161	5,355	6,320	4,315	10,635
45 - 64	5,198	1,643	6,841	3,008	1,067	4,075	8,206	2,710	10,916
65+	3,556	1,715	5,271	1,235	618	1,853	4,791	2,332	7,124
<b>TOTAL</b>	<b>12,500</b>	<b>6,092</b>	<b>18,592</b>	<b>8,030</b>	<b>4,376</b>	<b>12,406</b>	<b>20,530</b>	<b>10,468</b>	<b>30,998</b>

Source: (1)

Tables might also be used to present information on rates, ratios or proportions:

**Table 4**  
**Rate (per 100,000 Population) of Newly Diagnosed**  
**Cases of Tuberculosis by Age, Race and Sex, United States, 1973**

AGE	RATE (PER 100,000) BY RACE AND SEX						TOTAL
	White			Other			
	Male	Female	Total	Male	Female	Total	
0 - 4	5.2	4.4	4.8	22.0	19.4	20.7	7.4
5 - 14	1.5	1.7	1.6	9.5	8.7	9.1	2.8
15 - 24	3.9	3.7	3.8	21.9	22.9	33.4	6.4
25 - 44	11.1	6.8	8.9	90.8	44.5	65.7	15.9
45 - 64	28.0	8.1	17.7	148.9	45.9	93.7	25.3
65+	44.5	15.0	27.1	151.5	58.5	99.0	33.4
TOTAL	14.0	6.5	10.2	62.7	31.3	46.3	14.8

Source: (1)

Summarization of data--as appears in Tables 1-4--will be expedited and simplified by initially preparing a master table. In this master table, all available data should be tabulated in as great a detail as the available data permit. The advantage of such a table is that data relative to a single variable or to any combination of variables subsequently can be obtained for the preparation of summary tables without having to go back to the original data and retabulating it. Table 5 indicates a format of a master table from which Tables 1 through 4 could have been prepared.

Table 5

Number of Newly Diagnosed Cases of Tuberculosis, by Age,  
Race and Sex, United States, 1973

Age Group (years)	NUMBER OF CASES BY RACE AND SEX															
	White			Black			Indian			Oriental			Etc.	Total		
	M	F	T	M	F	T	M	F	T	M	F	T		M	F	T
< 1																
1 - 4																
5 - 9																
10 - 14																
.																
.																
.																
70+																
TOTAL																

At this point it should be re-emphasized that age often is the single most important variable in epidemiologic investigations, and consequently is one of the most frequently used variables for tabulating case data. Therefore it is appropriate to explain how to select age groups by which to tabulate various disease-specific data. There are two practical conditions which set limits on the age-groups selected as a basis for tabulating data: (1) the manner in which age is recorded in the original case reports or case data sheets (that is actual age vs. preselected age groups), and (2) the age groups by which denominator data are available, since ultimately, numbers usually are converted to rates. Within these limits, many possible combinations of age are possible, and the preferred age groups vary by disease and purpose of the analysis. In general, it is preferable to use age-groups that are widely used by other health agencies, in order to facilitate comparisons (see Table 6). Usually the age groups shown in Table 5 will be adequate for the master table, since from them any of the sets of age groups shown in Table 6 can be derived. The age groups shown in Table 6 are commonly used in CDC disease surveillance reports. They were selected originally as being the most likely to reveal meaningful differences in distributions of cases.

**Table 6**

**Some Age-Groupings Used to Tabulate Age  
Distributions of Cases of Disease**

Diphtheria, Hepatitis A&B, Meningococcal infections, Tetanus, Salmonellosis	Syphilis (P&S), Gonorrhea	Tuberculosis	Trichinosis, Leptospirosis, Encephalitis (arthropodborne)	Measles, Rubella
< 1 year				< 1 year
1 - 4		0 - 4 yrs.	0 - 9 yrs.	1 - 4
5 - 9	0 - 14 yrs.			5 - 9
10 - 14 or 10 - 19		5 - 14	10 - 19	10 - 19
15 - 19	15 - 19			15 - 19
20 - 24 or 20 - 29	20 - 24	15 - 24		
25 - 29	25 - 29		20 - 29	
30 - 39	30 - 39	25 - 44	30 - 39	20+
40 - 49	40 - 49		40 - 49	
50 - 59	50+	45 - 64	50 - 59	
60+			60 - 69	
		65+		
			70+	
<b>Total</b>	<b>Total</b>	<b>Total</b>	<b>Total</b>	<b>Total</b>

Sources: (2)

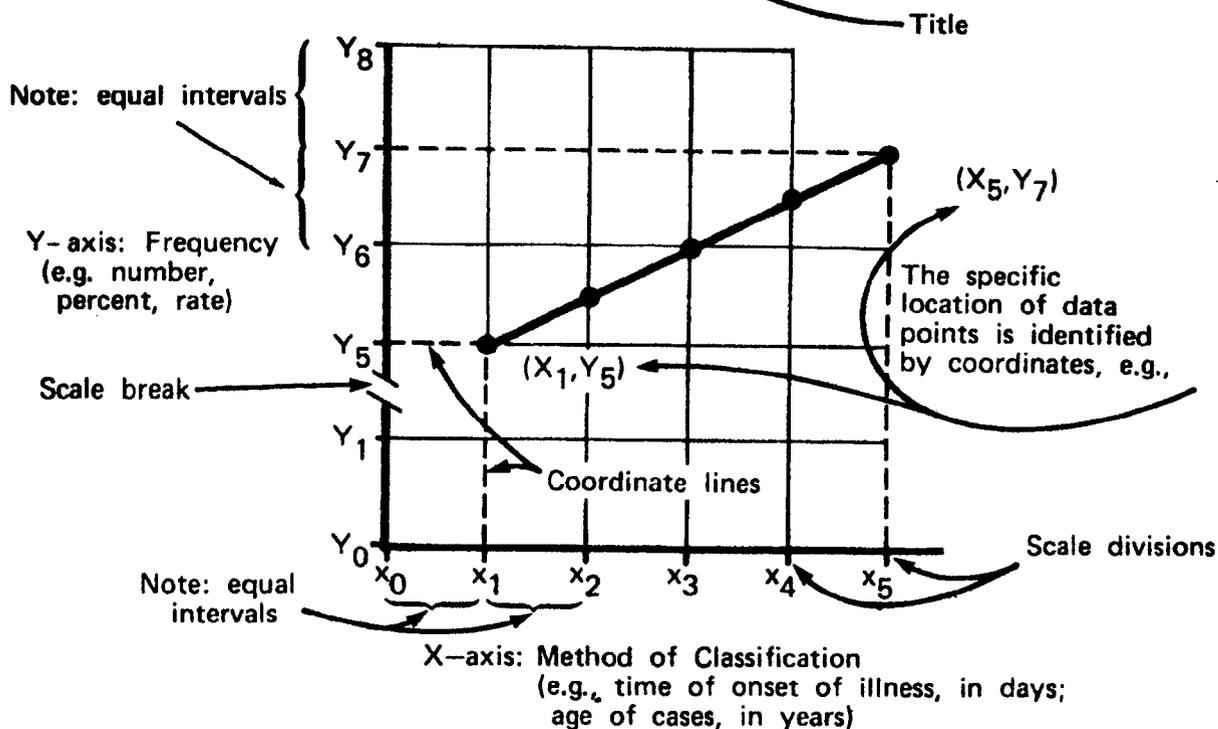
## GRAPHS

A graph is a method of showing quantitative data using a system of coordinates. There are several different types of graphs, such as rectangular coordinate, polar coordinate, and special purpose types (3 dimensional). This discussion is limited to rectangular coordinate graphs.

Rectangular coordinate graphs are those which consist of two sets of lines which intersect at right angles to each other. On each axis there is a scale of measurement and an identifying label. Figure 1 presents the general structure of rectangular coordinate graphs. Usually the variable assigned to the x (horizontal)-axis is the independent variable (method of classification, such as time or age) whereas the variable assigned to the y (vertical)-axis is the dependent variable (frequency of occurrence of an event). In a graph of real data, the symbols  $x_1, x_2, y_1, y_2$ , etc., would be replaced with appropriate sequential values of the data being portrayed.

Figure 1

A General Graph



Source: Adapted from (3)

Some of the most important points to remember in graphing are:

1. The simplest graphs are the most effective. No more lines or symbols should be used in a single graph than the eye can easily follow.
2. When more than one variable is shown on a graph, each should be clearly differentiated by means of legends or keys.
3. Every graph should contain enough information to be self-explanatory.
4. The title may be placed either at the top or bottom of the graph.
5. No more coordinate lines should be shown than are necessary to guide the eye.
6. Lines which outline the graph itself (usually the axes) should be drawn heavier than other coordinate lines.
7. Frequency is usually represented on the vertical scale and the method of classification on the horizontal scale.
8. Each axis in a graph may have a different scale, but the scales selected must be conducive to accurate interpretation of the data. On an arithmetic scale line graph equal intervals on an axis must represent equal numerical values.
9. Scale divisions on the axes should be clearly indicated. The label for each axis (e.g., the date of onset of illness of cases--on the x-axis) must specify the units (e.g., single days) into which the scale is divided.
10. A scale break may be used with a scale line graph, but, if used, it must be clearly identified as such.

#### Arithmetic Scale Line Graph

An arithmetic scale line graph is one in which equal distances along the y-axis represent equal quantities anywhere on that axis. Figures 2, 3, and 4 are examples of scale line graphs.

Note in these figures that the lengths of the x- and y-axes do not extend significantly beyond the values necessary to show the data contained in the graph. Note, too, that although each of the graphs shows a set of data having a different range of values, and each spans a different length of time, they can be adequately displayed in a relatively small space by proper selection of a scale. The selection of a scale for the y-axis usually involves:

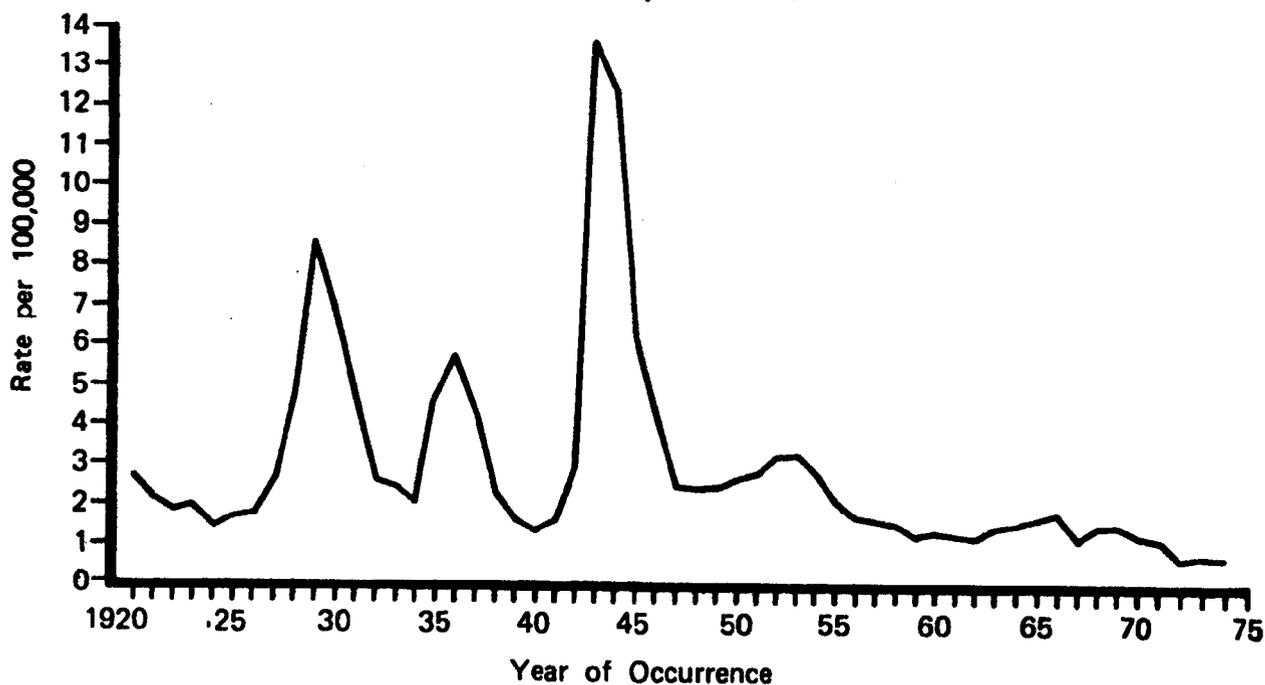
- (a) identifying the maximum, and subsequently the range, of the values to be shown on the axis--and rounding this figure off to some number slightly larger than the range to be portrayed;
- (b) counting the number of spaces, or squares, along the axis to be used which are considered necessary to show the data in sufficient detail for your purposes;
- (c) divide the number of squares available by the upwardly-rounded range of values to obtain the number of squares, or spaces, per unit value.

The scale or class intervals for the x-axis usually are those used in the table from which the data come. If the original data are too detailed for your purpose then some sub-groups or class intervals can be combined to obtain the desired larger groupings or class intervals.

When the scales for the x- and y-axis have been established, the graph should then be drawn centered in the page or other space to be used.

Figure 2

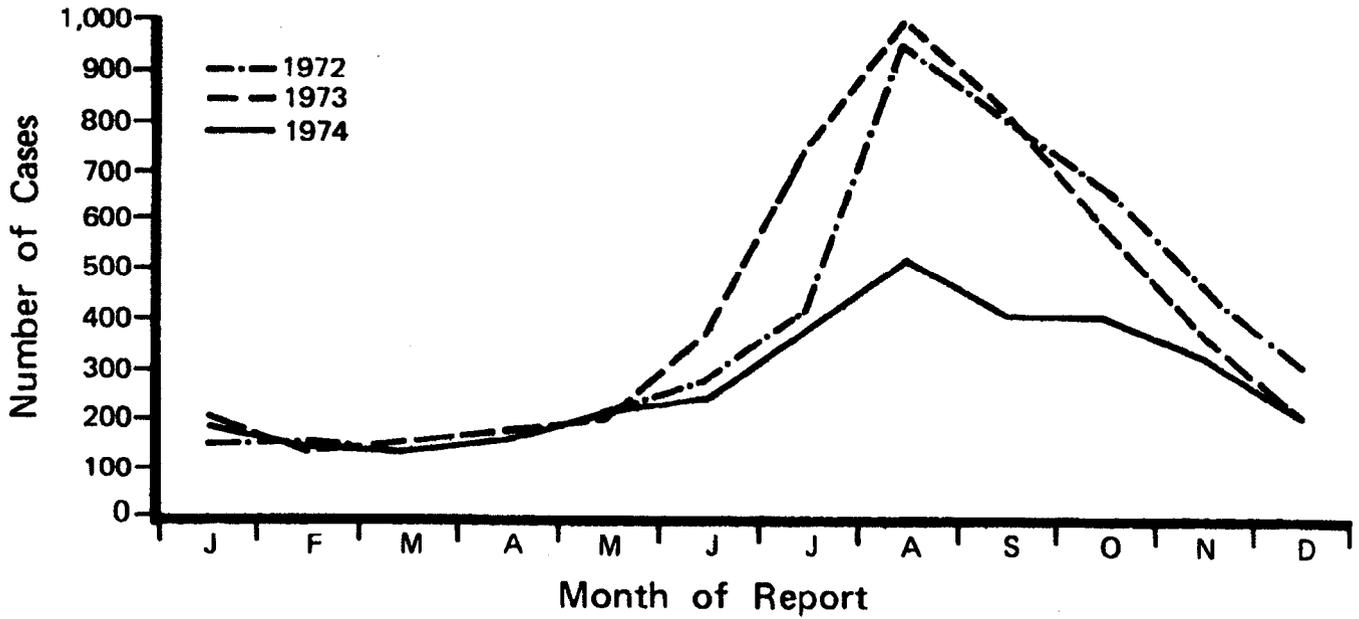
Reported Meningococcal Infections per 100,000 Population by Year of Occurrence, United States, 1920-1974



Source: (1)

Figure 3

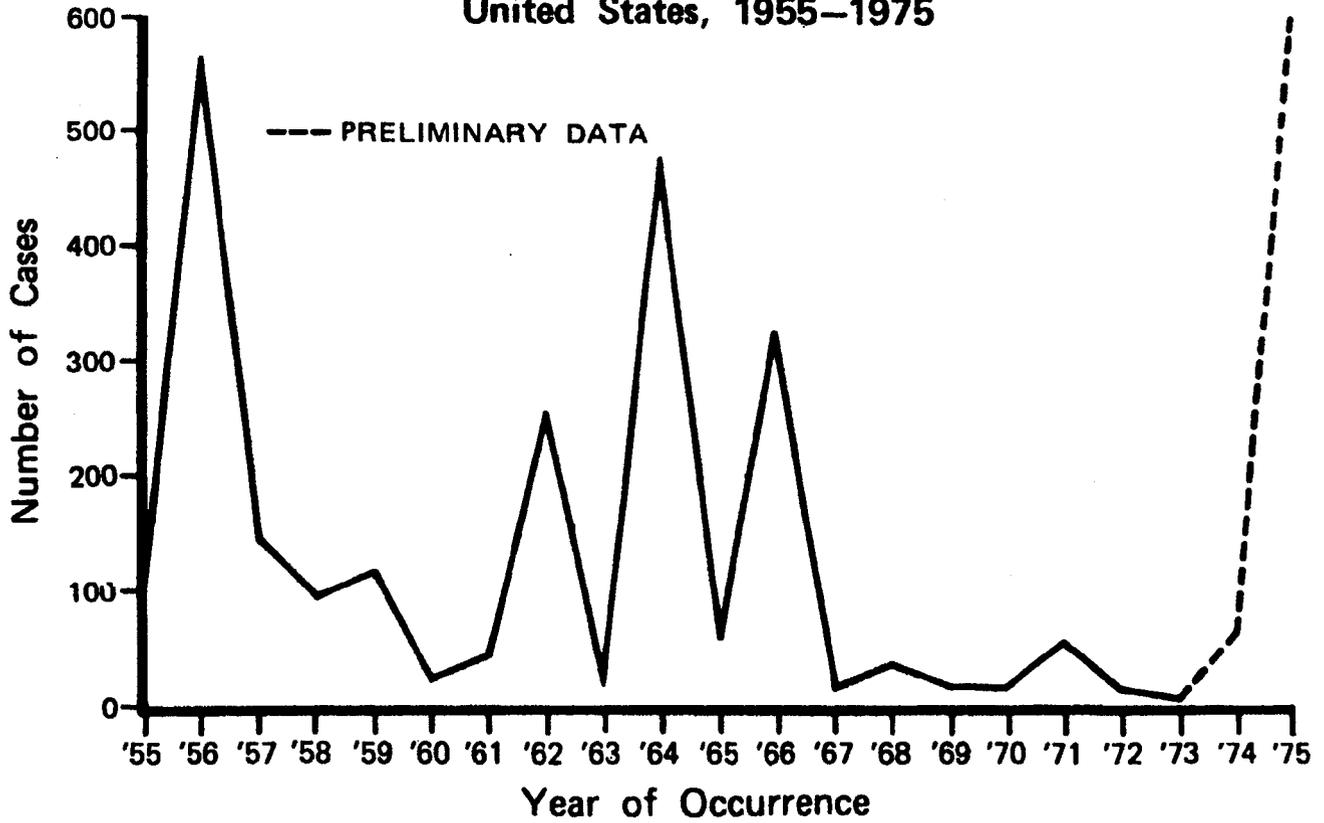
Reported Cases of Aseptic Meningitis, By Month of Report, United States, 1972-1974



Source: (1)

Figure 4

Reported Cases of St. Louis Encephalitis, by Year of Occurrence, United States, 1955-1975

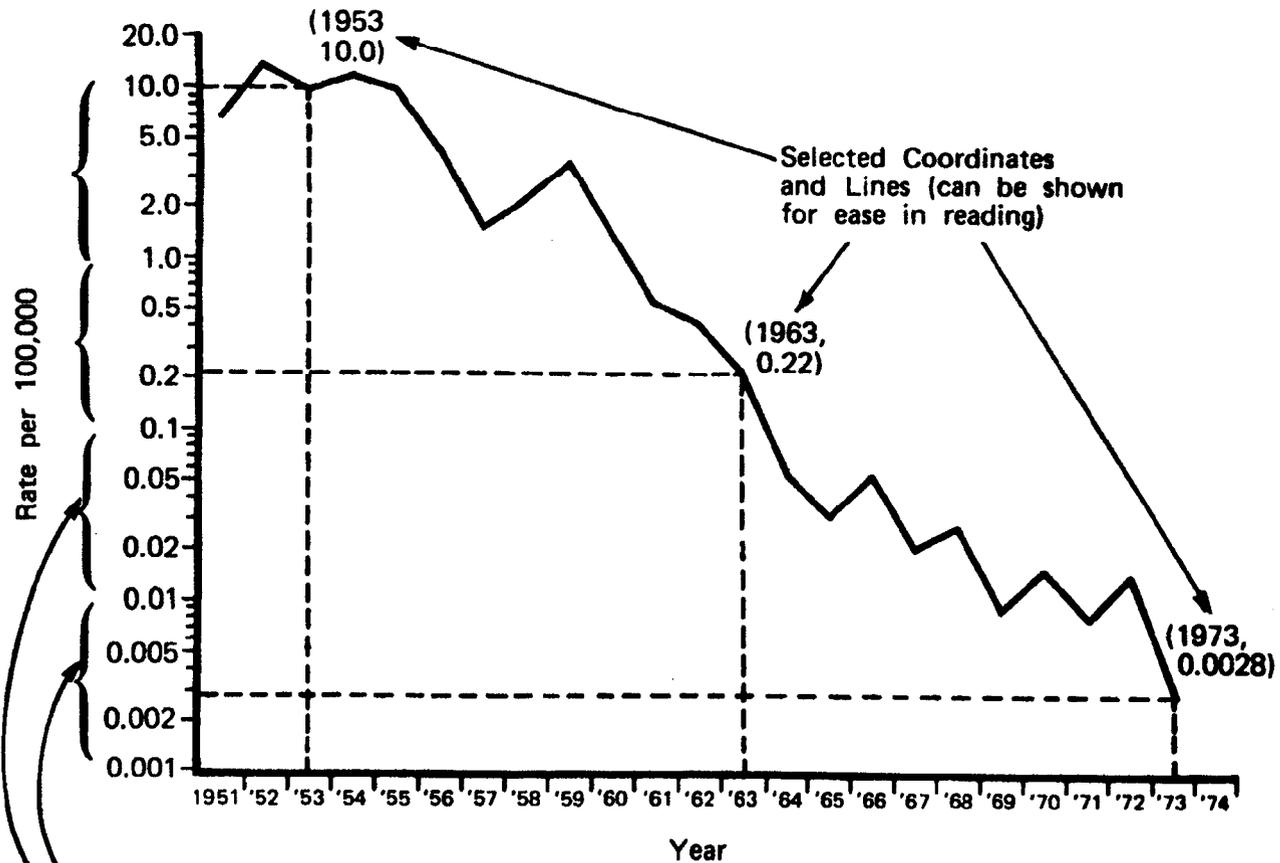


Source: (4)

Semi-logarithmic Scale Line Graph

The semi-logarithmic scale line graph is one in which the y-axis is measured in logarithms of units. (The x-axis is measured in arithmetic units, just as in the arithmetic scale line graph). This is useful when examining a series of data over a period of time and we are interested in the relative (or rate of) change in the values rather than the absolute magnitude of the values for the actual amount of their change. Figures 5, 6, and 7 are illustrations of this type of graph.

**Figure 5**  
**Reported Cases of Paralytic Poliomyelitis per 100,000 Population**  
**by Year of Occurrence, United States, 1951-1974**



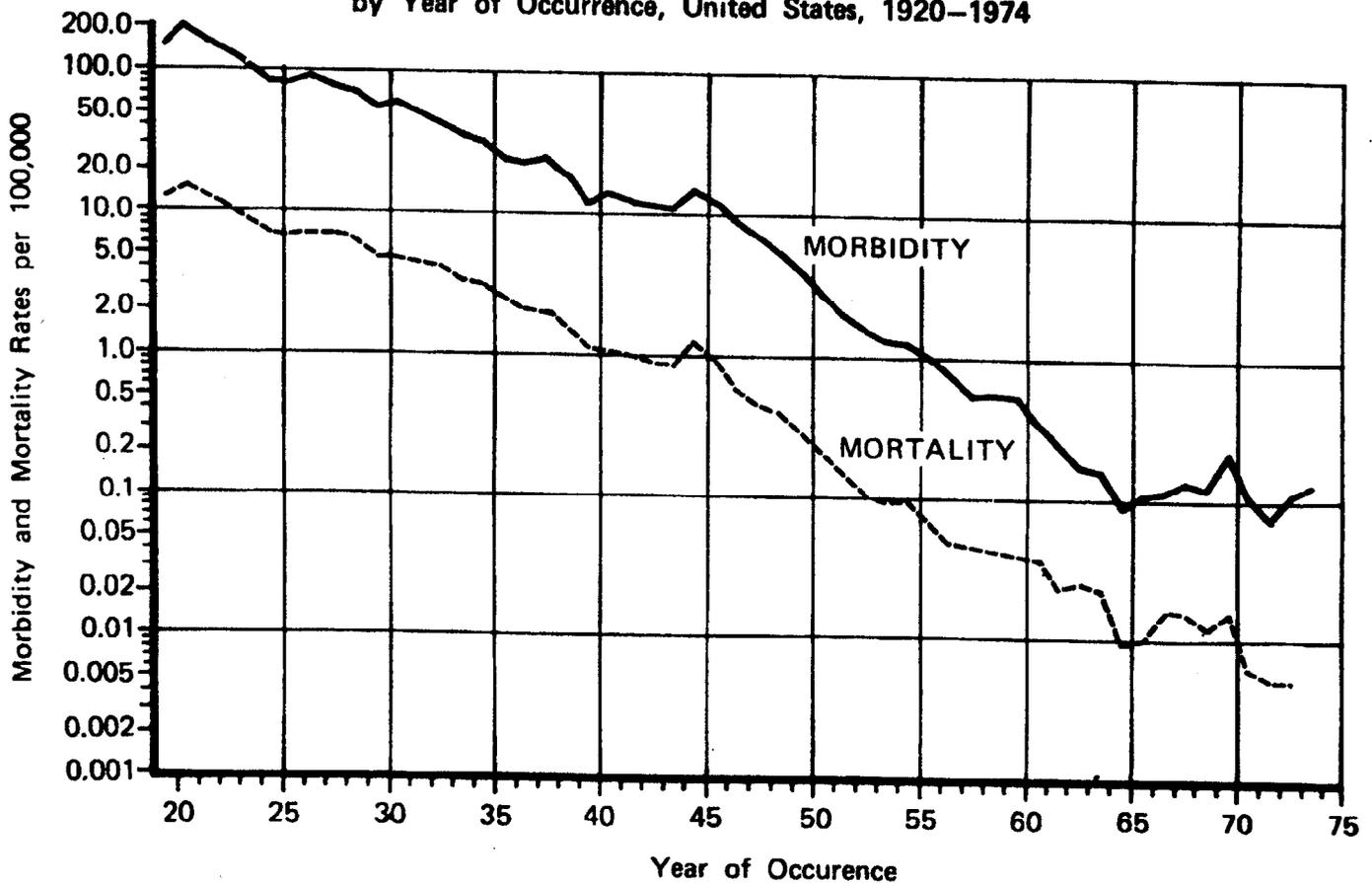
Source: (1)

Each cycle is the same length on the scale; and each cycle represents values ten times greater than the preceding cycle.

The advantages of semi-log graphs are:

- a. The slope of the line indicates the rate of increase or decrease.
- b. A straight line indicates a constant rate of increase or decrease in the values (or, if the line is horizontal, no change).
- c. Two or more lines following parallel paths show identical rates of change.
- d. Large changes or differences in the magnitude of numbers can easily be shown on a relatively small graph.

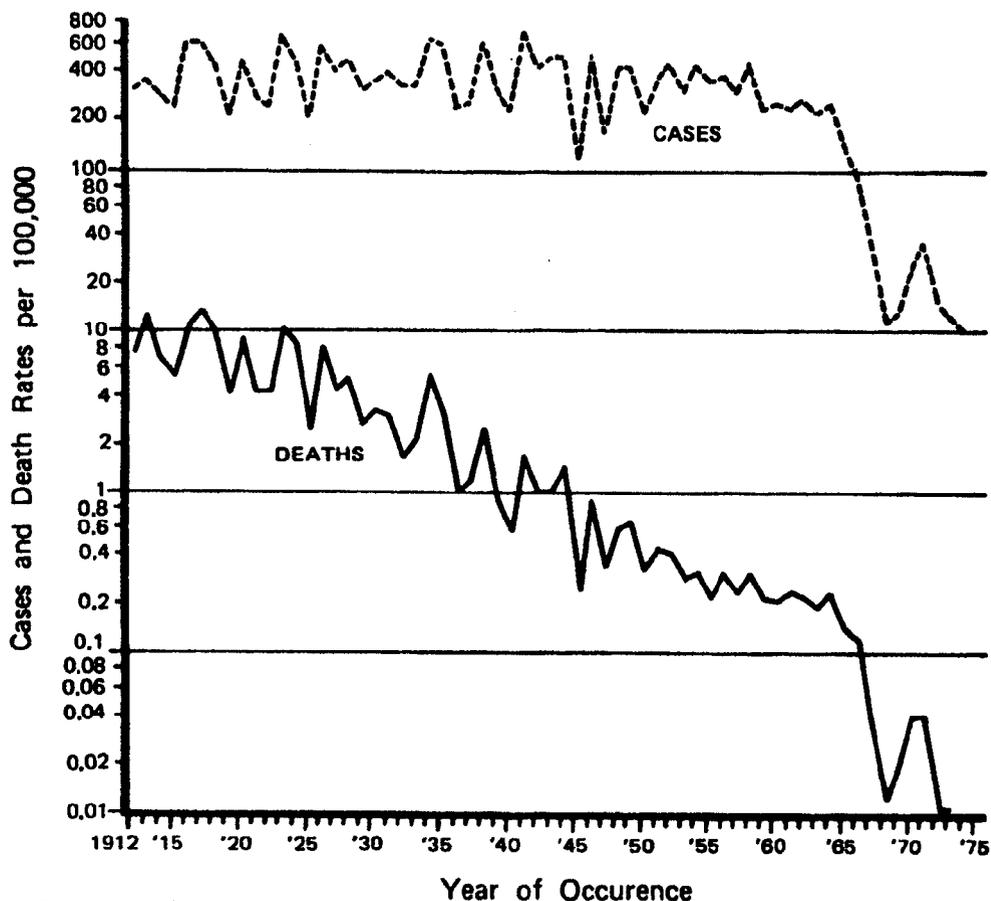
Figure 6  
Reported Cases and Deaths of Diphtheria per 100,000 Population  
by Year of Occurrence, United States, 1920-1974



Source: (1)

Figure 7

Reported Cases and Deaths of Rubella per 100,000 Population by  
Year of Occurrence, United States, 1912-1974



Source: (1)

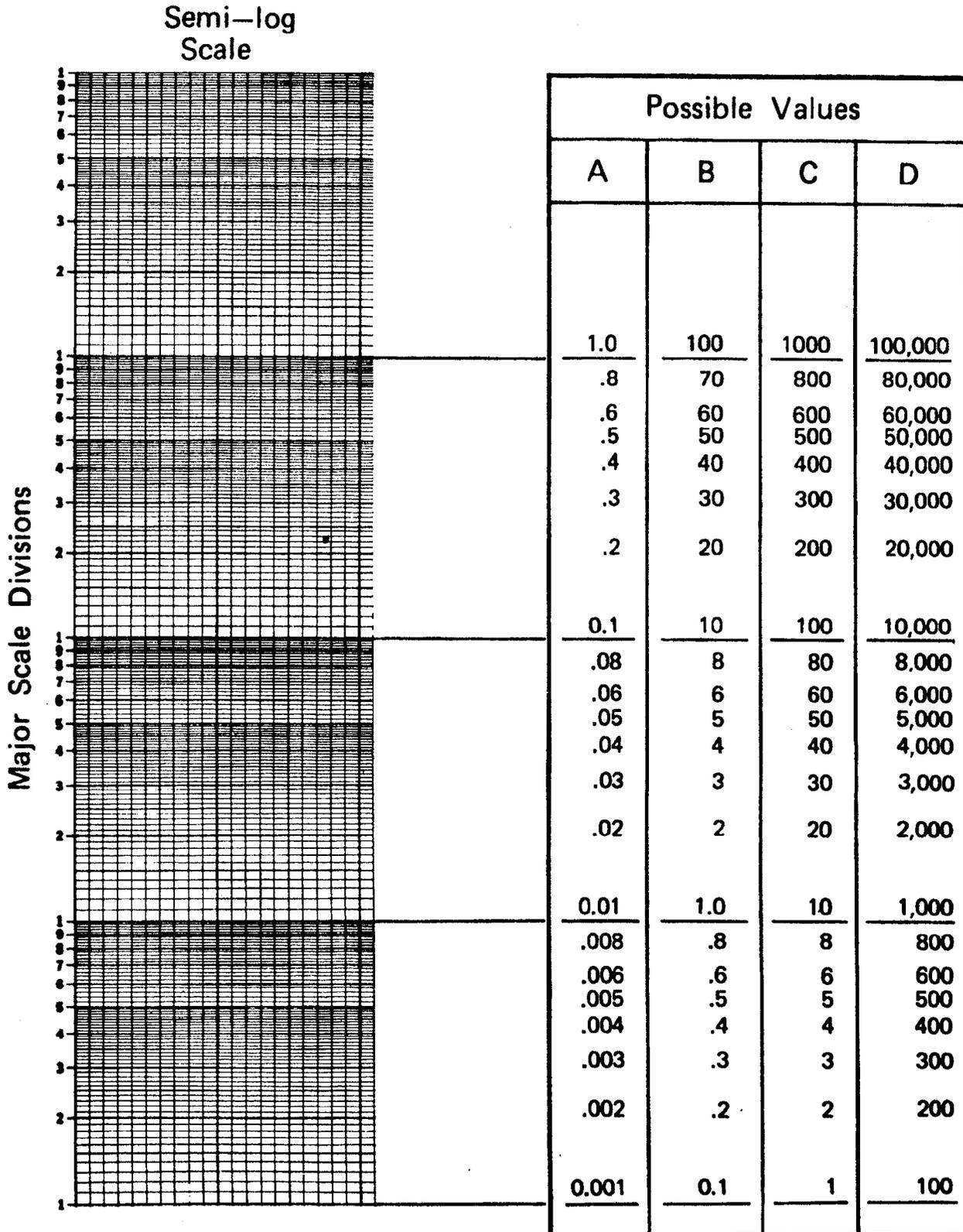
A point that merits special attention is the manner in which the values to use on the y-axis of a semi-logarithmic scale line graph are chosen and identified on the graph. You will notice in Figure 7 on the y-axis that the distances are all the same between 0.01 and 0.1, 0.1 and 1.0, and 1.0 and 10.0. It is characteristic of semi-logarithmic scales that any two values which differ by a multiple of 10 (that is, any two numbers the larger of which is 10 times greater than the smaller) are an equal distance apart on the scale. In Figure 7, therefore, the distances on the scale between 0.02 and 0.2, 0.01 and 0.1, and 0.05 and 0.5 are also all equal.

Referring then to the semi-logarithmic scale in Figure 8, the major scale division's (at the left side of the graph, marked with a 1) might represent-- at the user's discretion--any of the sample sets of values shown to the right of the scale and underlined, or any other set of multiples of ten. The specific set of values selected by the user would depend only on the range of values (minimum to maximum values) to be shown on the graph.

The choice, then, between using an arithmetic scale line graph and a semi-logarithmic scale line graph is made primarily on the basis of whether you want to show the absolute magnitudes of a set of values (arithmetic graph) or whether you wish to emphasize rates of change (semi-logarithmic graph). If your first choice is an arithmetic graph but the range in the magnitude of the values to be graphed is awkwardly large, you might then use the semi-logarithmic graph.

**Figure 8**

**An Illustration of Values Which Could be Assigned to The Y - Axis of a Semi-logarithmic Scale Line Graph**



**Histogram**

A histogram is a graph of a frequency distribution. Special features of histograms include the following:

- a. The width of the vertical bars is proportional to the width of the class intervals used; and,
- b. The height of the bars in a class interval is proportional to the frequency of occurrence of the event in that class interval. Because of this area characteristic, the easiest type of histogram to construct is one of equal class intervals as shown in Figures 9, 10, and 11. Also because of this characteristic the area of the rectangle in the legend must be exactly the same as the area of the same number of cases on the graph. Clearly, scale breaks cannot be used in histograms. Also, while the frequency distributions shown in the following examples are all based on numbers of cases, proportional distributions can also be displayed similarly.

**Figure 9**

**Hepatitis in Bone Marrow Transplant Patients and Platelet/Plasma Donors, Seattle LRC, 1972**

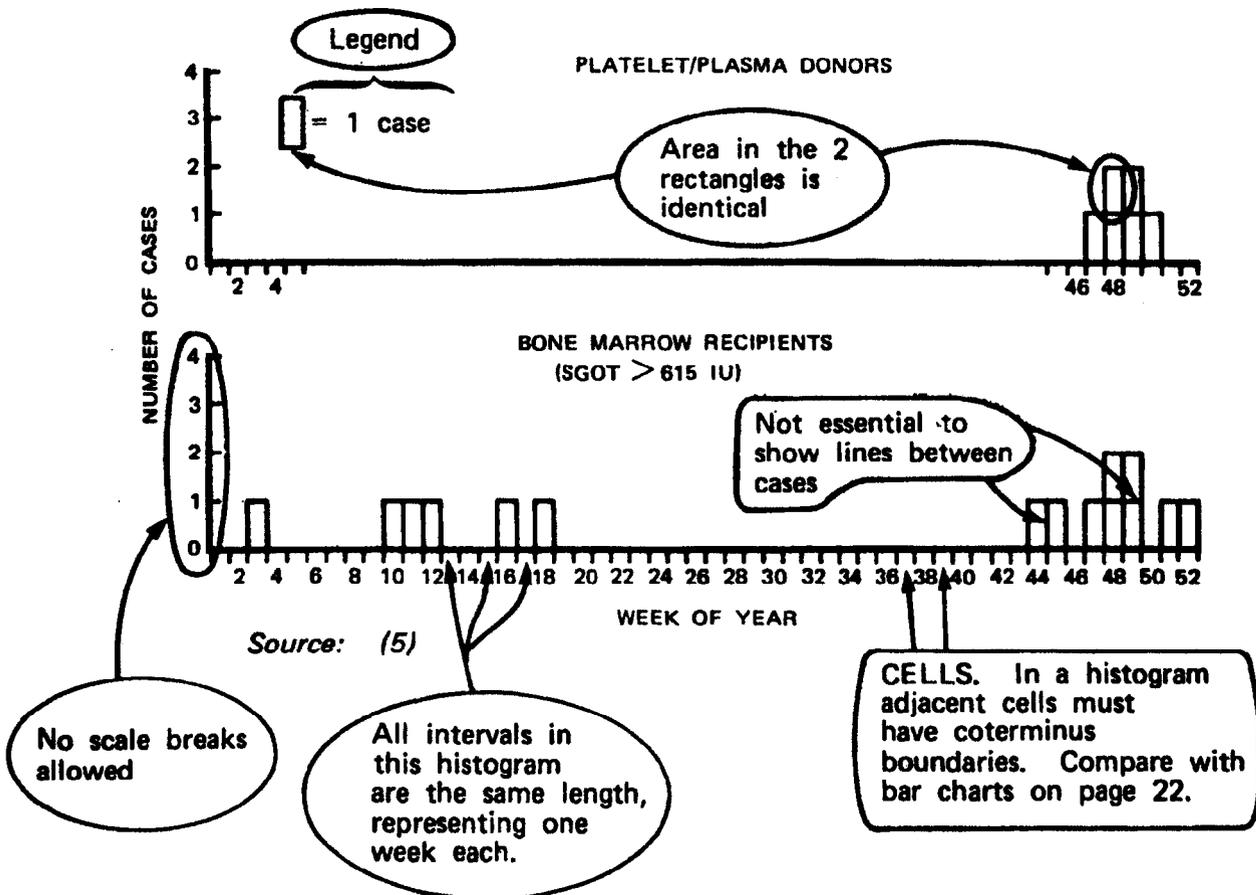
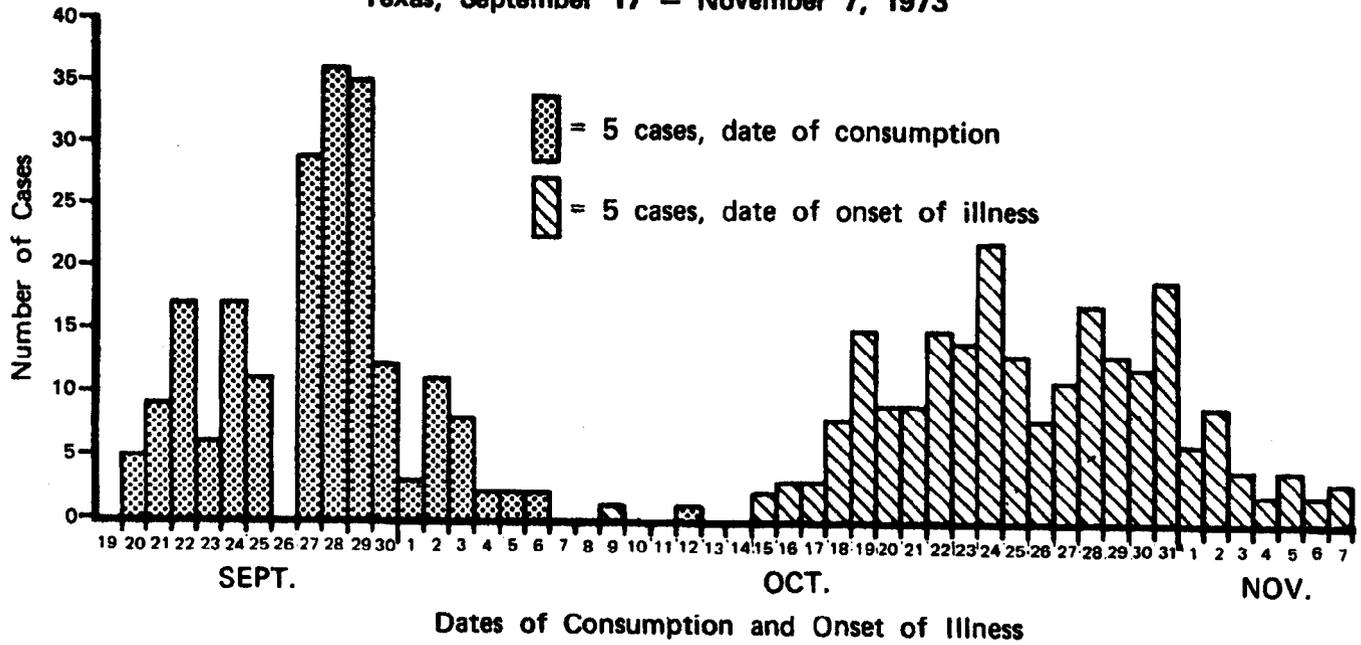


Figure 10

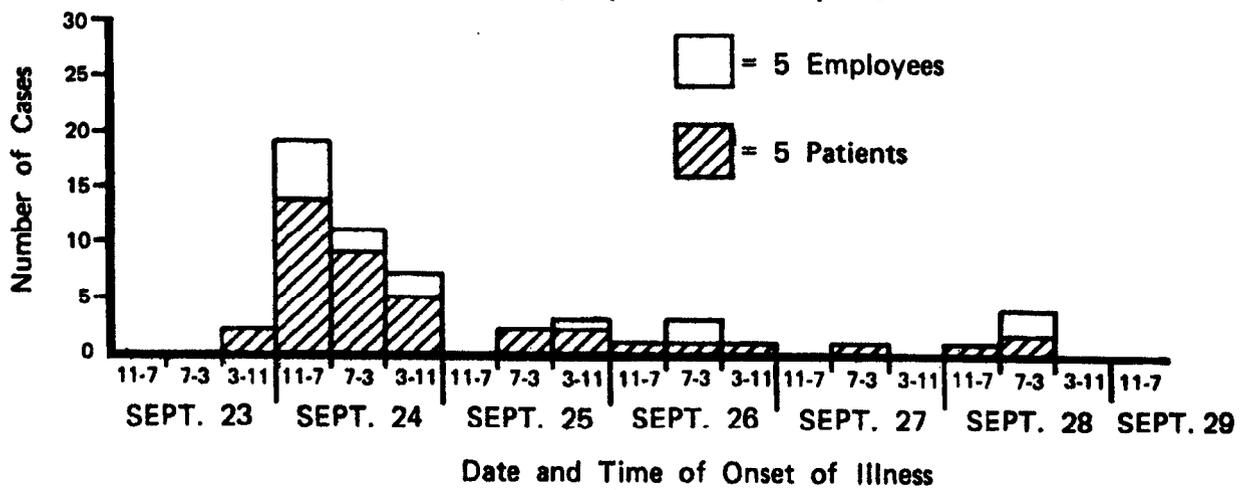
Cases of Oyster-Associated Hepatitis, by Date of Consumption and Date of Onset of Illness, Texas, September 17 - November 7, 1973



Source: (5)

Figure 11

Cases of Salmonellosis in Patients and Employees, by Date and Time of Onset of Illness, Pennsylvania Nursing Home, September 23-28, 1970



Source: (6)

## Frequency Polygon

If it is desired to present more than two sets of data in terms of a frequency distribution, the data are more clearly presented as a frequency polygon than as a histogram. A frequency polygon is constructed by plotting the individual values at the mid-point of their respective class interval and connecting them with a straight line as in Figure 12. In Figures 12 and 13 the frequency polygon was prepared from the same set of data as the corresponding histogram, which is included only to emphasize the similarities and differences between the two forms. Both enclose the same area. When making a frequency polygon, do not show the corresponding histogram on the same graph.

Figure 12

Number of Reported Cases of Influenza-Like Illness  
by Week of Onset, Sample City, 1970

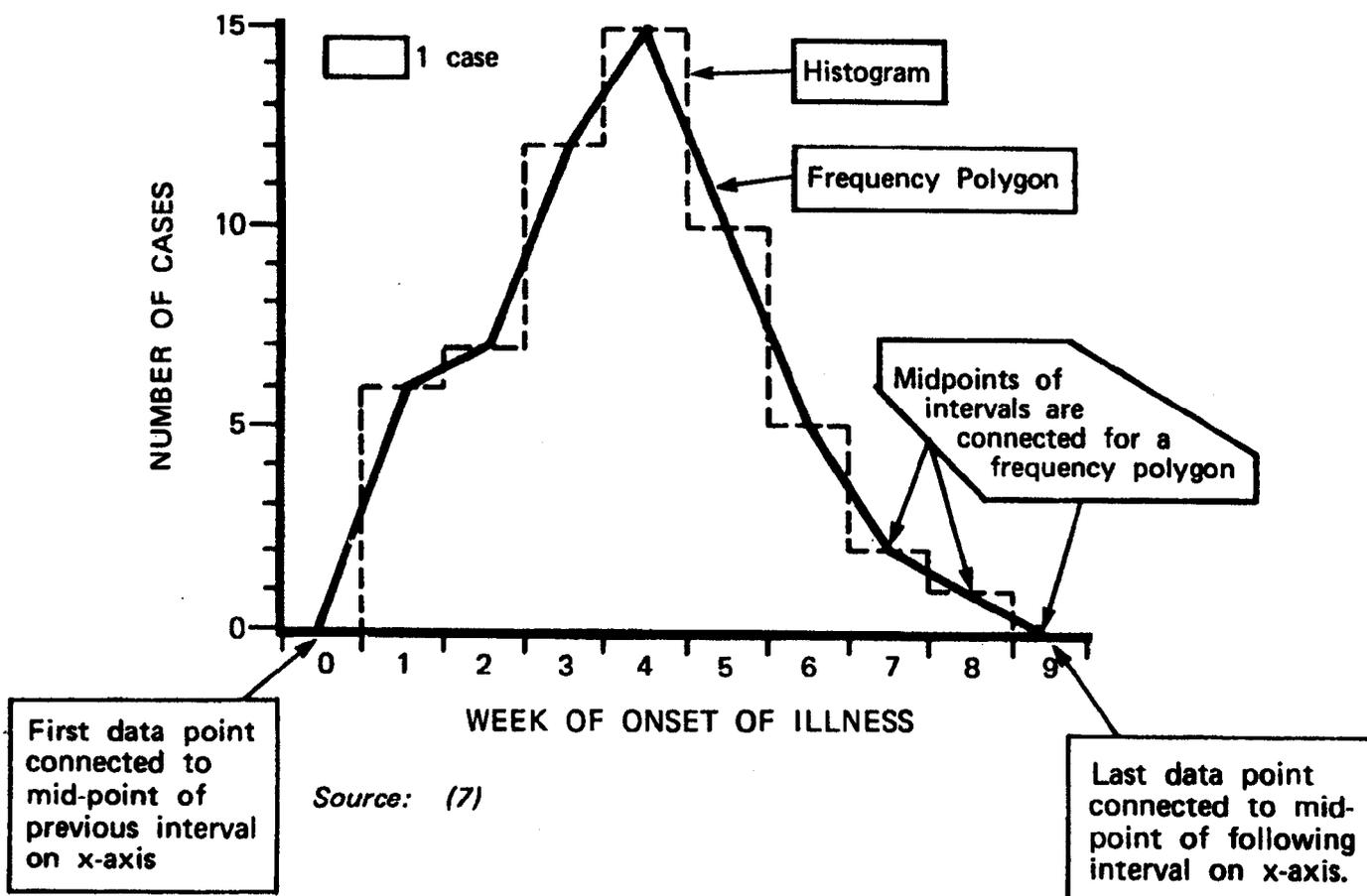
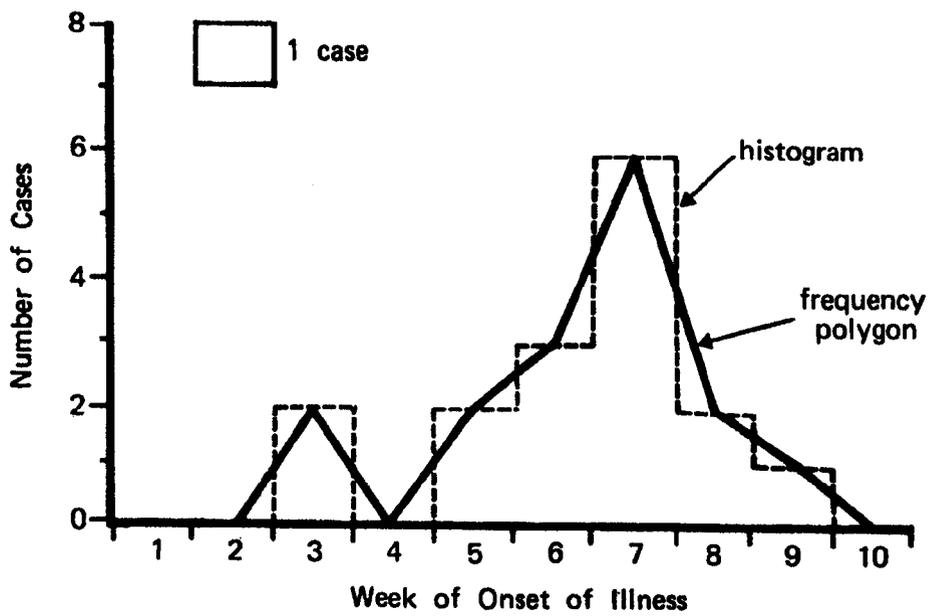


Figure 13

A Hypothetical Epidemic Curve, Sample City, 1970



Source: Case reports received by Sample City Health Department

The rules pertaining to the area under the curve apply to a frequency polygon just as they do to a histogram. The area in the frequency polygon must be equal to that which would have been in a histogram prepared from the same data. Therefore, it is necessary to "close" the frequency polygon. This closing is done by connecting the first and last data points with the base of the graph (the x-axis) at the mid-point of the preceding and following intervals, respectively. Figure 14 shows the correct method of closing the frequency polygon. Figure 15 shows an incorrect method.

Figure 14

**CORRECT METHOD OF CLOSING  
A FREQUENCY POLYGON**

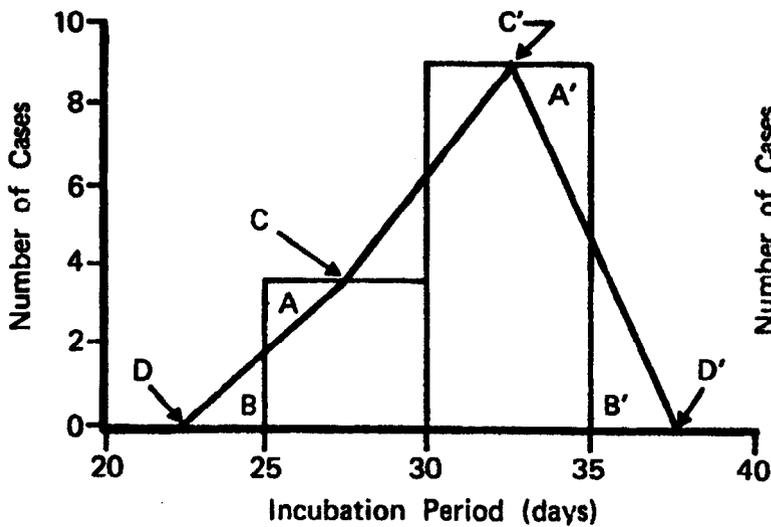
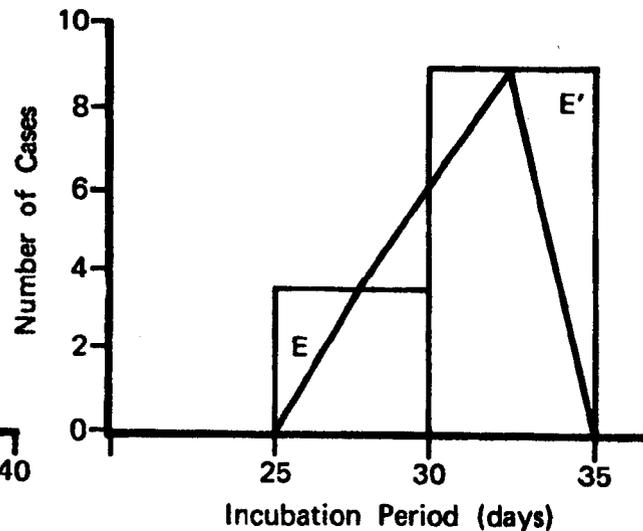


Figure 15

**INCORRECT METHOD OF CLOSING  
A FREQUENCY POLYGON**



Source: Adapted from (8)

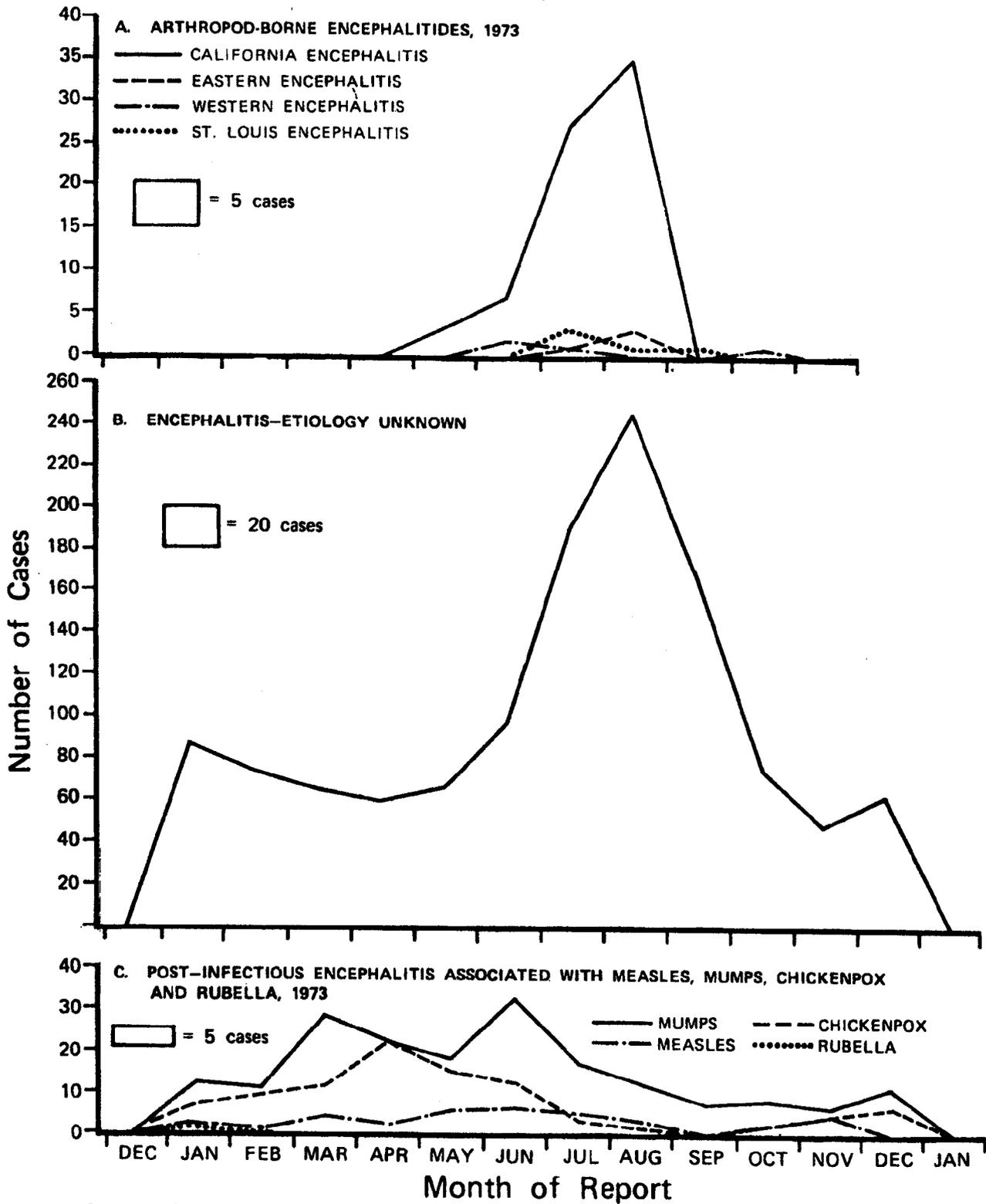
In Figure 14 the areas designated by A and A' would be part of the histogram if data were plotted by that method. In order to compensate for this area, which is excluded by the polygon, the point C is connected to the x-axis at D and C' to D', the mid-point of the preceding and terminal intervals, respectively, so that the areas designated by B and B' will be equal to areas A and A', respectively.

The method of closing a frequency polygon that is illustrated in Figure 15 is incorrect because the areas designated by E and E' are omitted and there is no provision for compensating areas.

Additional examples of frequency polygons are shown in Figure 16.

Figure 16

Reported Cases of Encephalitis by Etiology and Month of Report, United States, 1973



Source: (1)

## CHARTS

Charts are methods of illustrating statistical information using only one coordinate. They are especially appropriate for comparing the magnitudes of different events or of components of a total. There are many different types of charts, two of which are presented here: those based on length of a bar (a bar chart), and a geographical coordinate chart map.

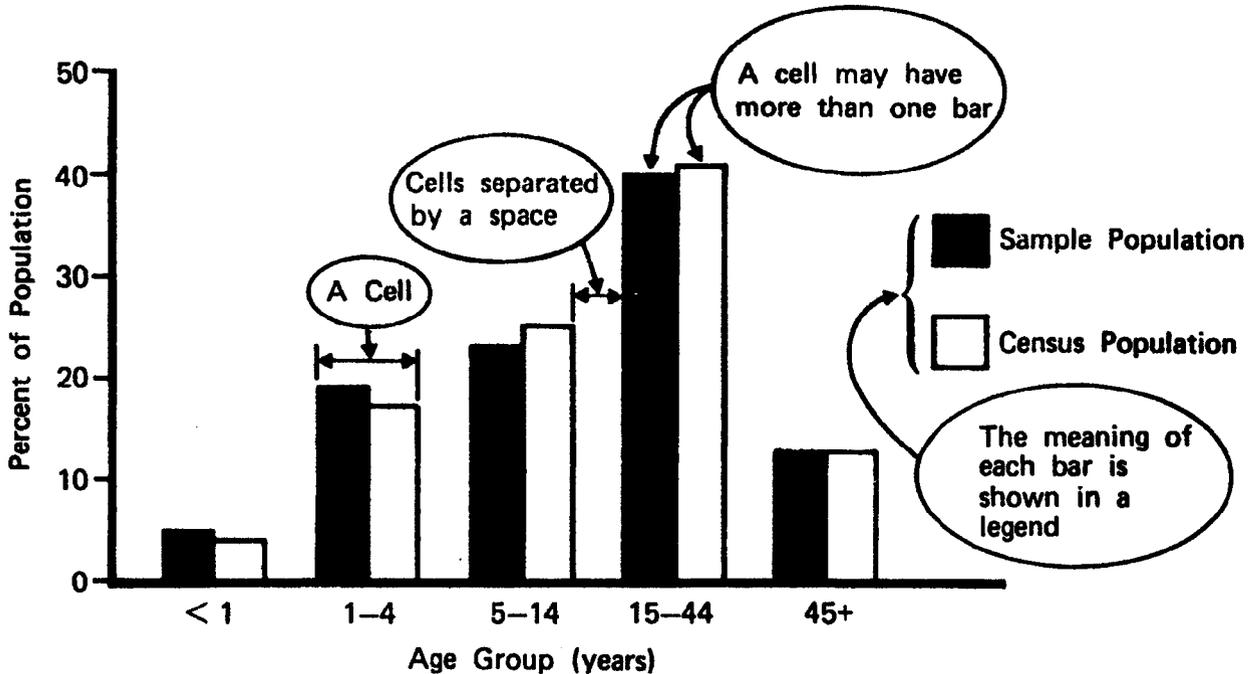
### Bar Chart

The bar chart has cells, all of the same column width (which is arbitrary), separated by spaces (see Figure 17). A cell may contain more than one bar; and if it does, the bars may be optionally separated with a space and must be illustrated distinctively. The distinctions must be shown in a legend. The length of each bar is proportional to the frequency of the event in the interval. The bars may be arranged in either ascending or descending order of height, or in some other arrangement selected to make a special point (e.g. Figure 17). They may be positioned either horizontally or vertically. A scale break should never be used with a bar chart as this could lead to misinterpretation.

While the main use of bar charts is to compare magnitudes (Figure 18), they can also be used to show frequency distributions (Figure 17) and time series data (Figure 19).

Figure 17

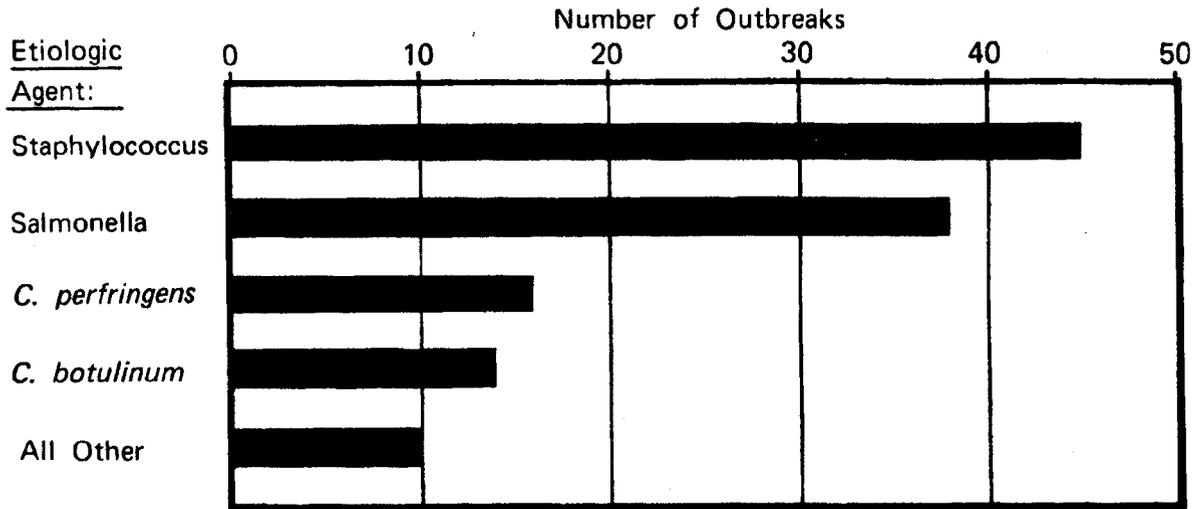
A Comparison of the Percent Distribution of the Age of The Population in a Sample Survey and The Population of The Whole Community, Oil City, May, 1975.



Source: Adapted from (7)

Figure 18

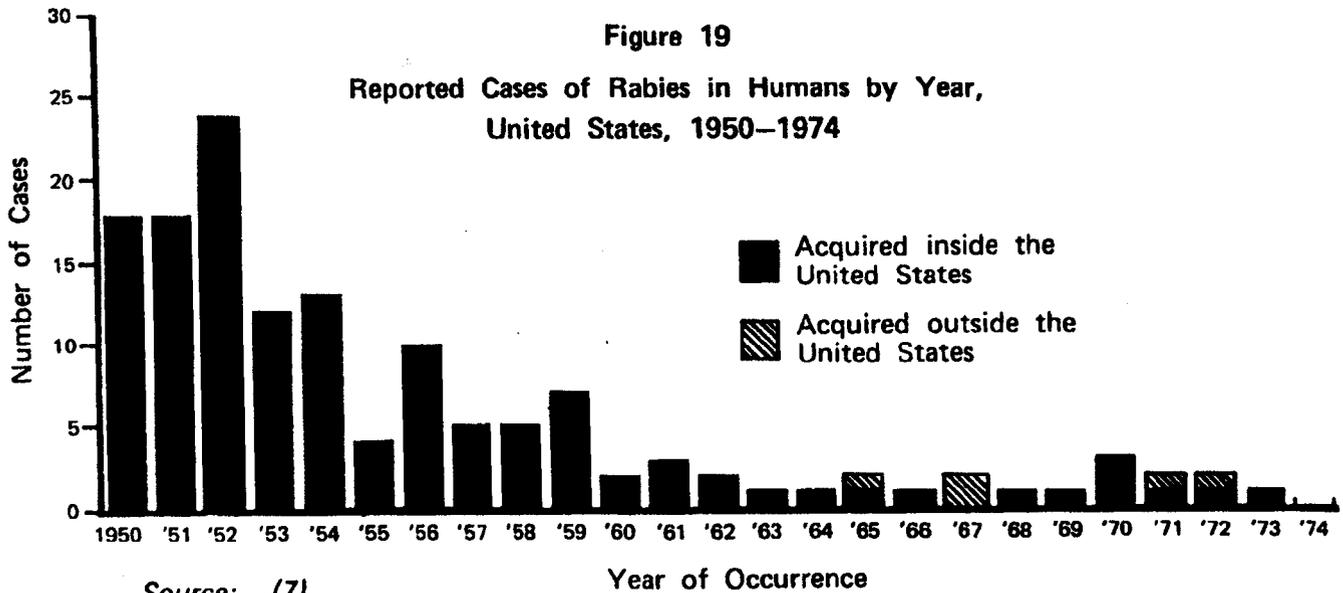
Number of Foodborne Disease Outbreaks of Known Bacterial Etiology Reported to the Center for Disease Control, by Etiology, United States, 1975



Source: Adapted from (2)

Figure 19

Reported Cases of Rabies in Humans by Year, United States, 1950-1974



Source: (7)

Geographic Coordinate Chart

Geographic coordinate charts are those charts which represent the occurrence of events using maps. Both spot maps and area maps are in common use. A spot map shows (Figure 20) by means of dots or other symbols, the location at which an event took place or at which a condition exists. An area map can show, by means of shaded or coded areas (Figure 21), either the incidence of an event in sub-areas or the areal distribution of some condition.

Figure 20

Reported Cases of Rocky Mountain Spotted Fever, United States, 1974

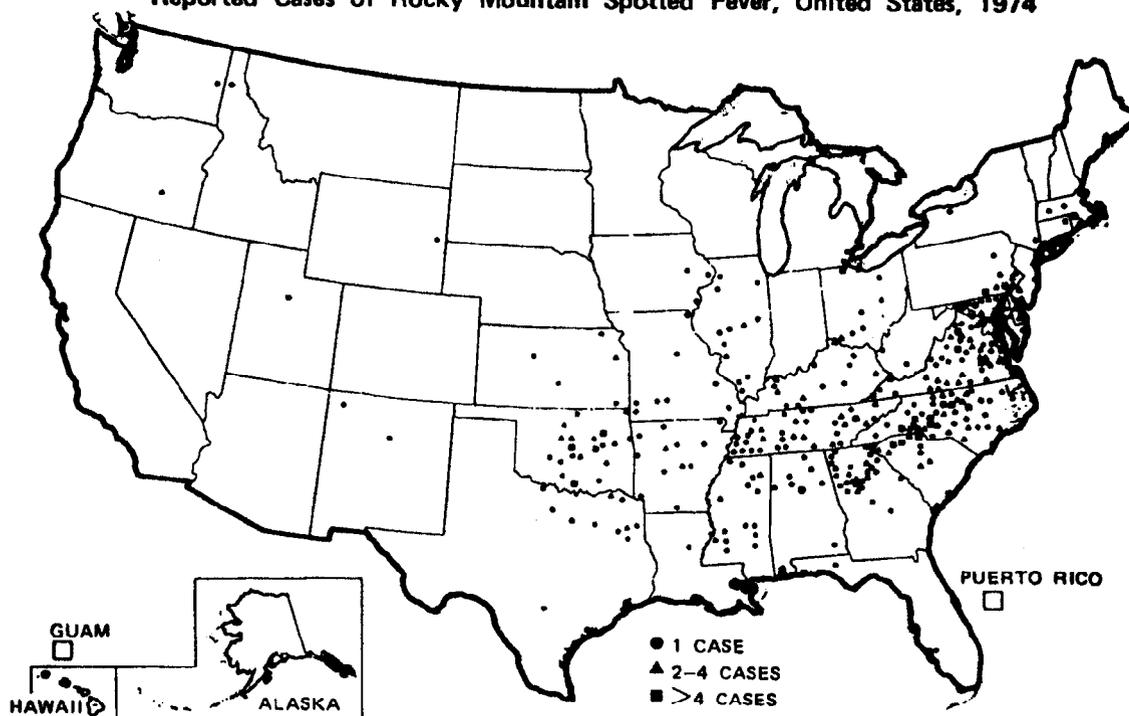
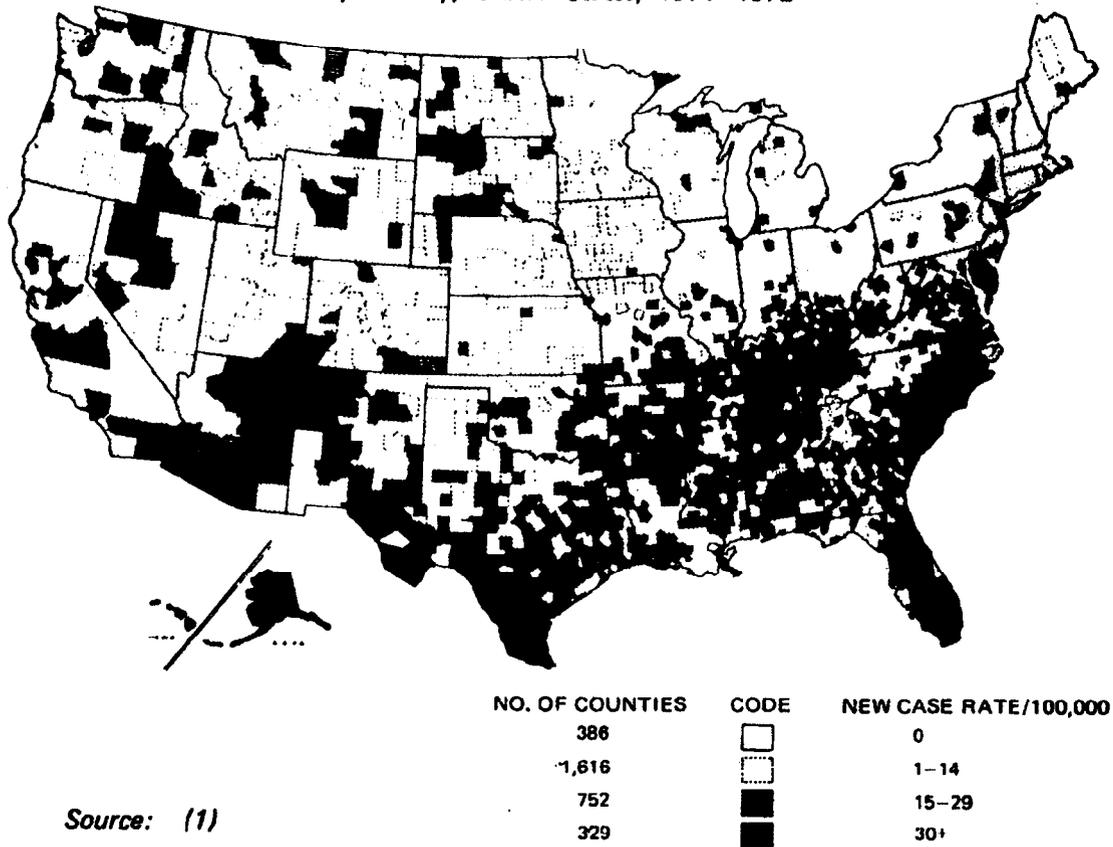


Figure 21

Reported New Cases of Tuberculosis per 100,000 Population by County, United States, 1971-1972



Source: (1)

Spot maps are made simply by placing a dot or other symbol on the map at the location on the map which represents the site at which the event occurred or the condition exists. If so many events occurred at one location that the dots lose their identity, then one or more additional symbols can be adopted to signify a range of frequency of events. Figure 20 is an example of this, with a legend that explains the symbols used.

The value of a spot map is in its portrayal of the geographic distribution of the occurrence of a class or kind of event. A spot map does not provide any measure of the risk of an event (e.g., of a resident acquiring a disease) occurring in a particular place--despite the number of dots that may be shown at that place--since the size of the population at risk of that event is not taken into consideration.

However, the area map shown in Figure 21 does identify the risk of an event occurring in an area since the population size is taken into account. That is, area-wide rates--rather than numbers--have been mapped.

Two important considerations in the construction of area maps--whether the areas are coded to represent numbers of events, rates, or something else--is how many different ranges of values to use (Figure 21 uses four, which are identified in the legend) and what the limits of each range shall be. In Figure 21 again, the four ranges were selected to show those areas in which events occurred: (a) not at all, in which the range is "0"; (b) at or below the national average, in which the range of rates is from greater than zero to less than 15/100,000; (c) from the national average to just less than twice as great as the national average, from 15 to less than 30 per 100/000; and (d) at a rate twice as great or greater than the national average, in which the range is 30/100,000 and greater.

Two other methods of selecting ranges of values are also commonly used. One of these (example A, page 27) is to rank-order all of the values to be coded and then arbitrarily divide the list of places into two, three, four, or more groups containing an equal number of members. If such a list of rank-ordered values were divided into four equal parts, for example, and the resulting four groups each assigned a separate code and mapped, the map would show the frequency quartile (highest, 2nd highest, 3rd highest, and lowest one-fourth) to which each coded geographic area belonged.

The other method (example B, page 28) is to identify the range from zero to the maximum value in the series to be mapped and then divide this range into some arbitrary number of class intervals of equal size. Codes then would be assigned to the intervals, and component areas mapped accordingly.

As an example of the use of these two methods, they are both applied in the following pages to the preparation of a map of the United States showing the newly reported cases of primary and secondary syphilis per 100,000 population by state in 1974. The first step using either method is to rank-order the rates as they are done in Table 7 on the next page.

Table 7

Rank-Order of Primary and Secondary Syphilis Rates per  
100,000 Population by State, United States, 1974

<u>Rank</u>	<u>State</u>	<u>Rate</u>	<u>Rank</u>	<u>State</u>	<u>Rate</u>
1.	SD	0.4	26.	CT	6.0
2.	MY	0.5	27.	CO	6.2
3.	VT	0.6	28.	AL	7.2
4.	WY	0.6	29.	PA	7.8
5.	NE	0.7	30.	KY	8.0
6.	UT	1.1	31.	NM	8.7
7.	ND	1.1	32.	MO	8.8
8.	WV	1.2	33.	IL	10.1
9.	IA	1.4	34.	NV	11.0
10.	NH	1.6	35.	MA	11.0
11.	ID	1.6	36.	TN	11.2
12.	RI	1.8	37.	NJ	11.5
13.	WI	2.2	38.	MS	11.7
14.	MN	2.2	39.	TX	11.8
15.	OH	3.1	40.	AZ	12.1
16.	AK	3.2	41.	DE	14.5
17.	IN	3.7	42.	VA	14.8
18.	KS	3.9	43.	LA	15.1
19.	WA	4.0	44.	NC	17.1
20.	HI	4.2	45.	MD	19.0
21.	ME	4.5	46.	CA	19.3
22.	AR	4.7	47.	NY	20.3
23.	MI	4.8	48.	GA	23.8
24.	OR	5.2	49.	SC	25.5
25.	OK	5.3	50.	FL	36.4

Source: (1)

A. Using the method of groups having equal numbers of members, we would proceed as follows:

1. Divide the list into four (or some other number) equal sized groups of places:

50 states  $\div$  4 = 12.5 members per group, resulting in the following groups:

- A. South Dakota through Rhode Island (1 thru 12)
- B. Wisconsin through Oklahoma (13 thru 25)
- C. Connecticut through Mississippi (26 thru 38)
- D. Texas through Florida (39 thru 50)

2. Identify the range of rates that correspond to the states that begin and end each subgroup:

<u>States</u>	<u>Range of Rates</u>
A. SD - RI:	0.4 - 1.8
B. WI - OK:	2.2 - 5.3
C. CT - MS:	6.0 - 11.7
D. TX - FL:	11.8 - 36.4

3. Adjust the ranges of rates so that there are no gaps between the end of one class interval and the beginning of the next:

<u>States</u>	<u>Original range of rates</u>	<u>Adjustment of limits of class intervals</u>	<u>Adjusted range of rates</u>
A. SD - RI:	0.4 - 1.8	$\longrightarrow 1.8 + \frac{1}{2}(2.2-1.8) = 2.0$	0.0 - 2.00
B. WI - OK:	2.2 - 5.3	$\longrightarrow 5.3 + \frac{1}{2}(6.0-5.3) = 5.65$	2.01 - 5.65
C. CT - MS:	6.0 - 11.7	$\longrightarrow 11.7 + \frac{1}{2}(11.8-11.7) = 11.75$	5.66 - 11.75
D. TX - FL:	11.8 - 36.4		11.76 - 36.40

4. Assign a symbol or area code to each range of rates and code each state using the code appropriate for that state's rate:

<u>States</u>	<u>Range of Rates (per 100,000)</u>	<u>Codes</u>
A. SD - RI	0.4 - 2.00	
B. WI - OK	2.01 - 5.65	
C. CT - MS	5.66 - 11.75	
D. TX - FL	11.76 - 36.4	

B. Using the method of class intervals of equal size, we would proceed as follows:

1. Divide the range of rates into four (or any other small number of) ranges of equal length:

$$\frac{36.4 - 0.0}{4} = \frac{36.4}{4} = 9.1$$

2. Add the result, 9.1, to the lowest rate, 0.0, four times to identify the four ranges of rates:
  - A. 0.0 through  $(0.0 + 9.10) = 0.0$  through 9.10
  - B. 9.11 through  $(0.0 + 9.10 + 9.10) = 9.11$  through 18.20
  - C. 18.21 through  $(0.0 + 9.10 + 9.10 + 9.10) = 18.21$  through 27.30
  - D. 27.31 through  $(0.0 + 9.10 + 9.10 + 9.10 + 9.10) = 27.31$  through 36.40
3. Assign a code to each range of rates and map each state according to its rate:

<u>Range of Rates (per 100,000)</u>	<u>Code</u>	<u>State</u>
0.0 - 9.10		SD - MO
9.11 - 18.20		IL - NC
18.21 - 27.30		MD - SC
27.31 - 36.40		FL

If a set of data were being mapped in which the number of cases or rate in some sub-areas was zero, then it would usually be preferable to establish a separate code for this group of sub-areas. The above ranges of rates, and area codes, would become:

<u>Range of Rates (per 100,000)</u>	<u>Code</u>
0	
0.01 - 9.10	
9.11 - 18.20	
18.21 - 27.30	
27.31 - 36.40	

**GLOSSARY OF TERMS**

## GLOSSARY OF TERMS

- Cells ..... Spaces in a table, graph, or chart in which data are entered.
- Class interval ..... For any given epidemiologic variable, the class intervals are the specific component sub-groupings of values. (e.g., for the variable "age," two class intervals could be less than 15 years old and 15 years old and older).
- Data (continuous) ..... Data consisting of measurements of things for which there are an infinite number of possible values between the minimum and maximum value in that set (such as age, weight, temperature and chemical concentration).
- Data (discrete) ..... Data consisting of measurements of things which can only be counted or measured in whole units (such as the number of persons meeting some criteria or other, the number of cells or organisms per unit volume, etc.).
- Frequency distribution ..... A tabulation of the number of times an event occurs in each class interval.
- Order (of magnitude) ..... A multiple of 10. Two values which differ in size by a multiple of ten are said to differ by an order or magnitude. If the values differed by a multiple of 100 (that is  $10 \times 10$ ) they would be said to differ by two orders of magnitude.