Vol. 1 No. 3

SCHS STATISTICAL PRIMER

State Center For Health Statistics
Division of Health Services
Department of Human Resources
P.O. Box 2091, Raleigh, N.C. 27602

Charles J. Rothwell, SCHS Director
George C. Robertson, **PRIMER** Editor

GRAPHS AND DATA DISPLAYS

Most of you think you already know this subject. You probably had enough classroom experience with graphs to feel comfortable with the simpler ones. But do you have a broad understanding of the various techniques, and do you always look critically at graphs? If not, this Primer is for you: its purposes are to review the sorts of graphs used most commonly in public health and to sound a warning about their misuses.

Graphs are pictures of data. Pictures are easier to grasp than tables of numbers, so for the sake of clear communication we try where possible to turn our data into pictures. Sometimes the results become cultural traditions. The bell-shaped, normal curve from statistics is a model that parents use to think about the heights of five-year-olds or their scores on a test. And news magazines traditionally report government spending using circular pie charts cut into slices that represent the amounts of money.

While some pictures of data are simple enough, others are based on complicated theory and may require quite a bit of effort to appreciate. And some graphs, as we shall see, are difficult because they misrepresent data. But wherever graphs are well presented they have enormous appeal, and no amount of words or intellectual effort will quite catch up with their intuitive worth. A graph is the sort of thing you cannot completely learn from definition: you must rather get to know it as you get to know a smell or a taste, the "atmosphere" of a small town, or the personality of an individual.

Arithmetic and Logarithmic Graphs

Typical graphs show values of a variable Y plotted against values of another variable X. Several kinds of graphs are illustrated in the figures. The familiar arithmetic graph, e.g., Figure 1, is the simplest. Regular calibrations on the axes (the X and Y reference lines) define a constant unit distance all over the graph. So a unit of response in the Y variable, no matter where it occurs or what its direction, always has the same physical size on the paper. Most of us got to know the arithmetic graph in high school algebra.

Widely used and very important, logarithmic graphs are distinquished by axes with logarithmic scaling. A common variation is the semilogarithmic graph, exemplified by Figure 2 which actually shows two plots on one set of axes. Figure 2 is semilogarithmic because only the vertical Y axis is scaled logarithmically; the horizontal X axis has ordinary arithmetic scaling. The obvious feature of a logarithmic axis is that distances become increasingly compressed: thus, in Figure 2, the nine units from 1 to 10 cover the same physical distance on the Y axis as the 90 units from 10 to 100, or as the 900 units from 100 to 1,000. A unit of response clearly shrinks in physical size as

Figure 1

CASES OF DISEASE A, BY MONTH

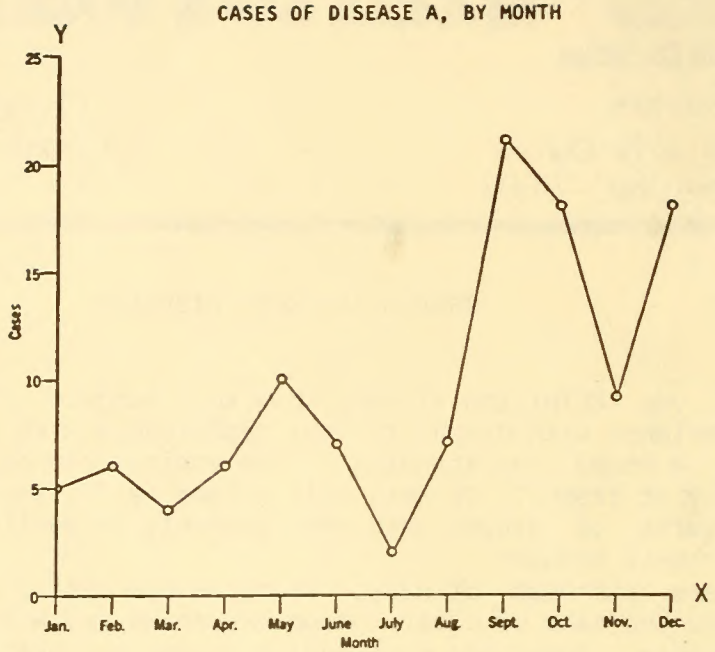
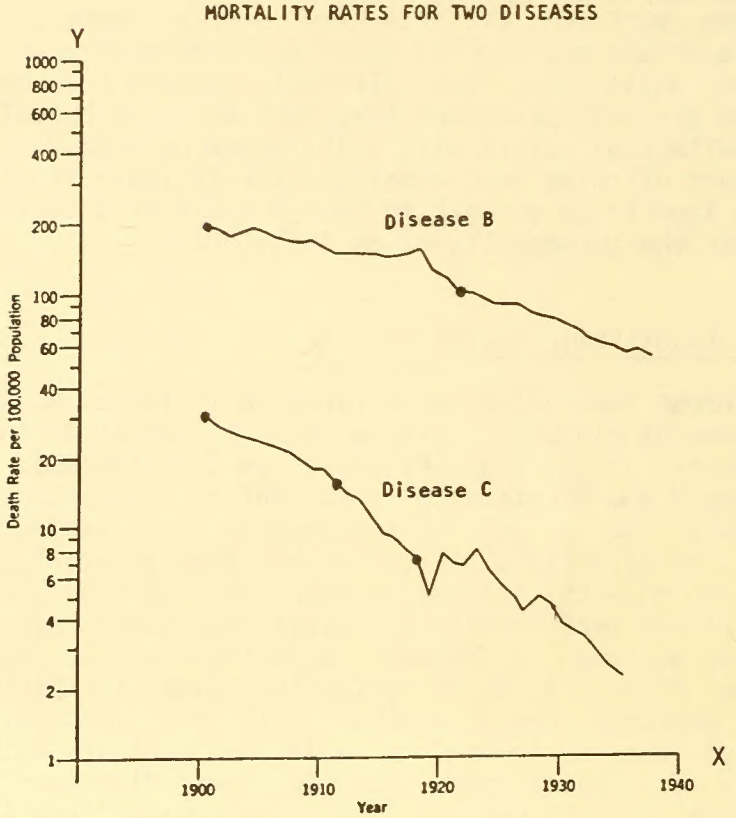


Figure 2

MORTALITY RATES FOR TWO DISEASES



it moves along the logarithmic axis. Two other features of logarithmic graphs should be mentioned. First, there is no zero on a logarithmic scale, so logarithmic variables cannot be plotted at the zero level. Second, any straight line in a logarithmic, or semilogarithmic, graph represents a geometric progression, meaning numbers like 2,4,8,16,32, . . . that increase by a constant proportion. Any curve that does not plot as a straight line does not represent a geometric progression. (We will explain this point below.)

Logarithmic graphs are important for two reasons:

1. They conveniently present data that vary over a wide range. Figure 2 does precisely this: the death rate for disease C has a few values as small as 3, and the death rate for disease B has values as large as 200. With an arithmetic Y axis, we would face a dilemma in the choice of scale. If we should calibrate the Y axis for the large rates of disease B, then the small rates of disease C would nearly vanish into obscurity because they would fall into a thin band along the bottom of the chart. But if we should calibrate the Y axis for the small rates, then the Y axis would have to be extremely long to accommodate the large rates; indeed, a vast empty space would appear on the graph between the two plots. A logarithmic Y axis lets us avoid the dilemma. Logarithmic ruling spreads apart small values, so that small rates for disease C become distinct; yet large rates for disease B can also be satisfactorily plotted without requiring the Y axis to have an unwieldy length.

2. A logarithmic graph shows proportional, or percentage, changes in a variable. Neither the arithmetic amount of change nor the variable's value at any point is of major importance in typical logarithmic graphs. To explain how proportional changes can be deduced from a logarithmic graph we must go back to the point made above: geometric progressions, with constant percentage changes, plot as straight lines on a logarithmic scale. A straight line results because the logarithms of a geometric progression form an arithmetic progression. For example, the common logarithms of 10, 100, 1,000, 10,000, . . . are, respectively, 1, 2, 3, 4, . . ., an arithmetic sequence which forms a straight line on a regular arithmetic scale. And the common logarithms of 2, 4, 8, 16, 32, . . . likewise form another arithmetic sequence which would plot as a straight line. It is actually logarithms that are being plotted in logarithmic graphs; that is, the plot of actual data points on a logarithmic scale is equivalent to a plot of corresponding logarithms on an arithmetic scale. And that is why a logarithmic axis becomes increasingly compressed.

The conclusion from these ideas—we must hold it in mind always when looking at logarithmic plots—is that equal distances on a logarithmic scale represent equal percentage changes, whereas equal distances on an arithmetic scale represent equal numerical changes. And a corollary of this conclusion is that equal slopes (or degrees of slant) on two logarithmic plots indicate equal rates of percentage change. Applying these generalizations in Figure 2, we see that for disease C the mortality rate dropped 50% between 1900 and 1912, then halved itself again between 1912 and 1918. (Three dots are placed on the plot for disease C so that the reader may project them to the axes and verify the preceding sentence.) For disease B the mortality rate dropped 50% between 1900 and about 1922; but as of 1938, when the study of disease B ended, the halving had not been repeated. (The two dots on the plot for disease B should be projected to the axes exactly as were the three dots for disease C.) Comparing the general downward slopes of the two plots, we see that the rate of percentage change in mortality from disease B is less than the rate of percentage change in mortality from disease C.

Figure 3
URBAN AND RURAL POPULATION
OF THE U.S., 1790-1960

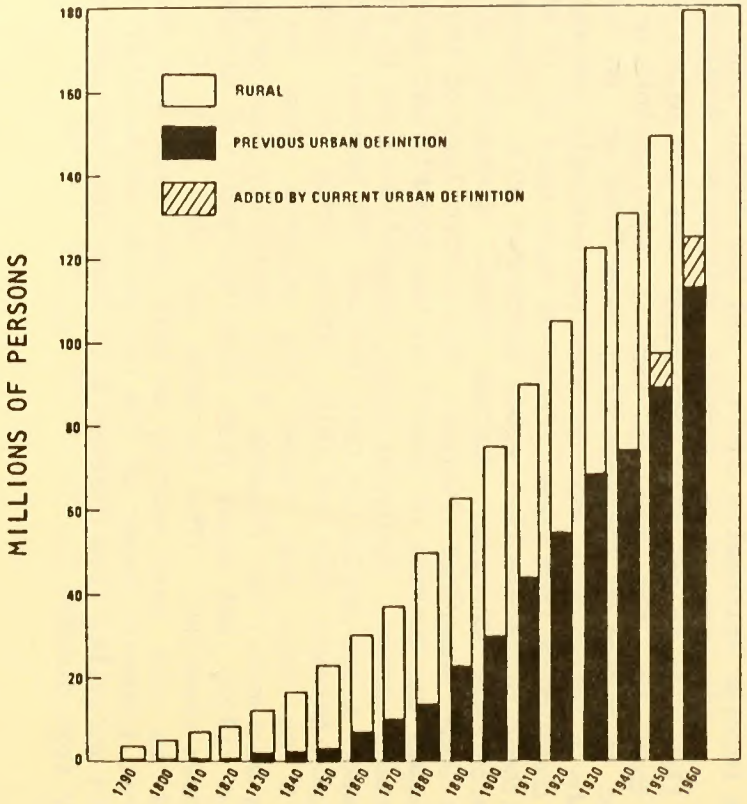


Figure 4
EXAMPLE OF POPULATION PYRAMID

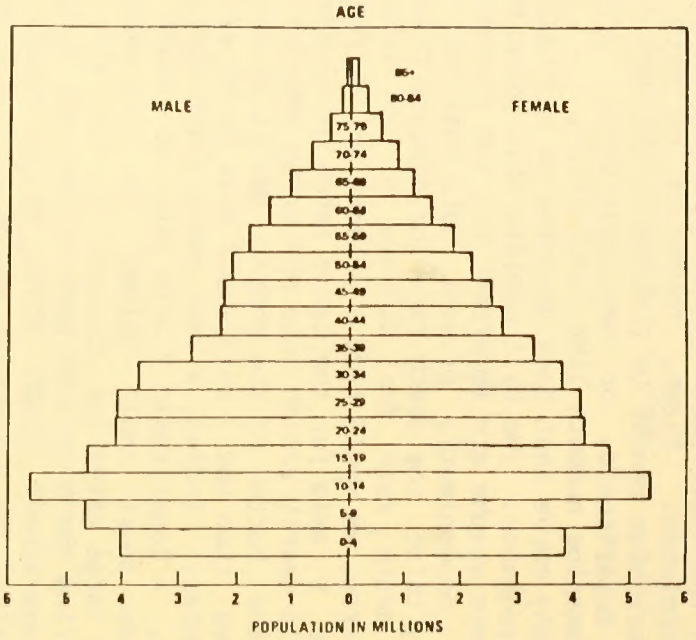


Figure 5

POPULATION CHANGE
APRIL 1970 TO JULY 1979

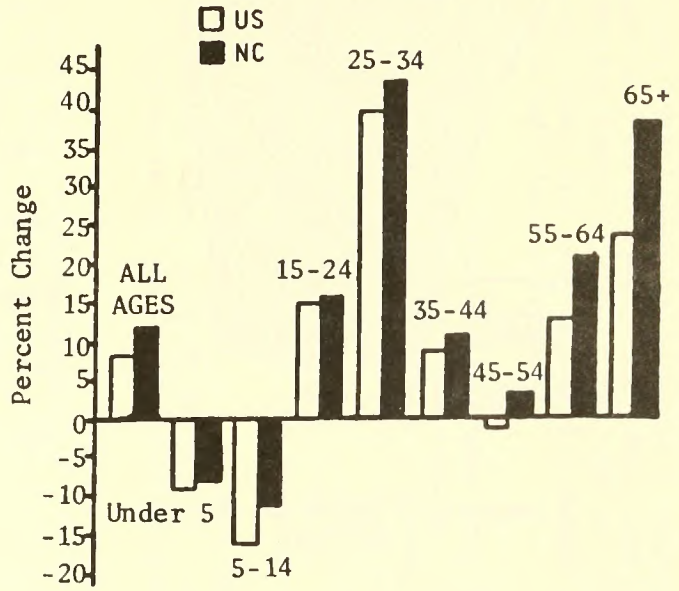


Figure 6
DISTRIBUTION BY COUNTRY OF
BIRTH OF POPULATION X

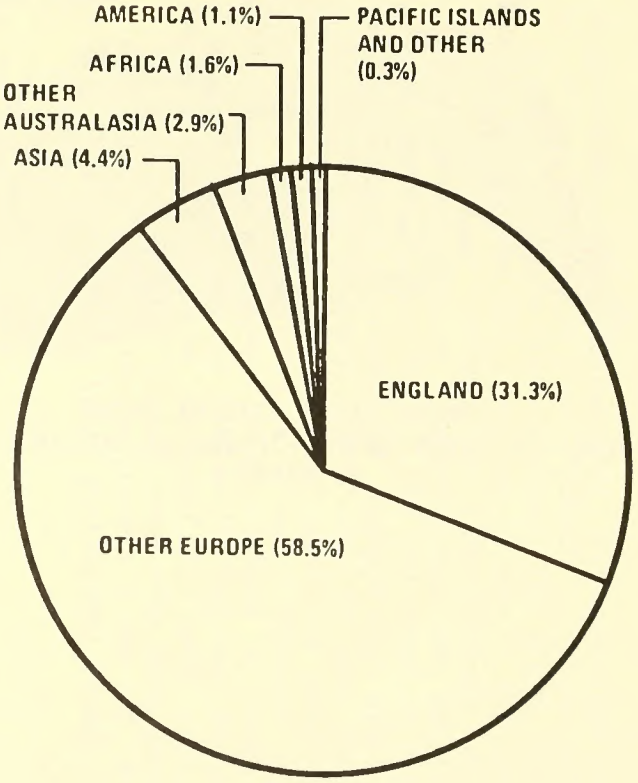


Figure 7
DEATHS BY AGE AND SEX FROM DISEASE D

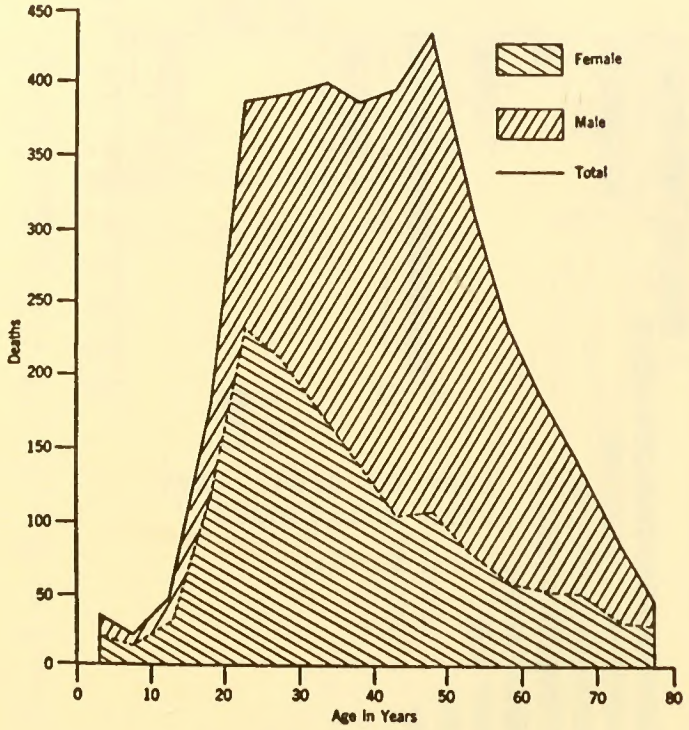
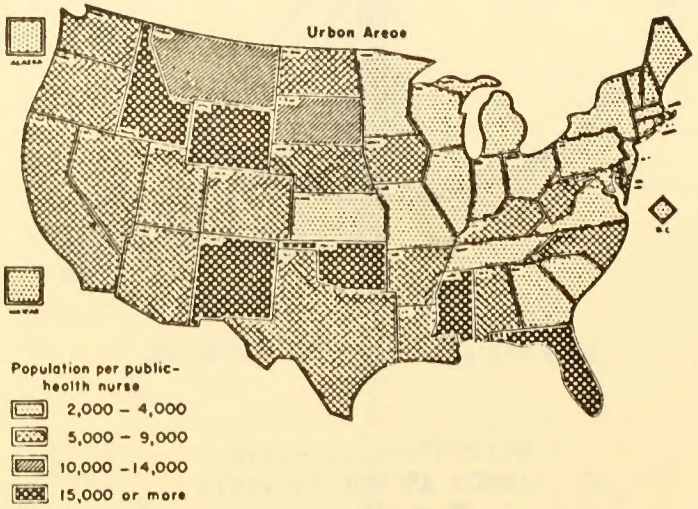


Figure 8
POPULATION PER PUBLIC-HEALTH NURSE IN URBAN AREAS IN EACH STATE, 1940



Other Kinds of Data Displays

Figure 3 shows a bar graph of population versus year, with each bar divided to show three subpopulations. Note that rural and "added" figures must be obtained by subtraction. Figure 4 shows a "double-sided" bar graph, the kind universally used by demographers to show age distributions by sex. (Histograms, frequency distributions, and cumulative probability distributions will be discussed in a future Statistical Primer.) Figure 5 demonstrates how bar graphs can display negative as well as positive quantities. The pie chart in Figure 6 shows a population broken into subsets by slices of appropriate size. The striking arithmetic plots in Figure 7 almost shout at us that there are differences by sex; note that numbers of male deaths must be obtained by subtraction. (See also the discussion of Figure 7 under "Perils" below .) Figures 8 and 9 are conformant maps. A conformant map depicts a study area subdivided into a number of data zones (e.g., the nation divided into states, or the state divided into counties). The data value assigned to each data zone translates into symbolic shading. To simplify such maps, the data zones are usually grouped into 4 to 6 clusters, each of which represents a set of similar data values. (Clustering will be discussed in a later Statistical Primer.)

The Perils of Bad Graphs and Charts

We speak of "hard data" and say, "Numbers don't lie," as though there is something innately truthful about data. Maybe there is. But the way numbers are made into graphs can misrepresent the truth.

There are two kinds of "bad" graphs. In the first, the data are badly presented through ignorance or oversight on the part of the author. We deplore the problem but usually can correct it by vigilant reading.

The second kind of "bad" graph is, alas, another matter. This kind is purposefully concocted to deceive. The intent is to shock a reader, stampede him into action, or render him purblind to an issue. All that data users can do to defend themselves is to have foreknowledge of the tricks and to approach graphs with a bit of healthy skepticism. Here are just a few of the most common tricks. Much of the terminology comes from Huff (see reference).

1. The Gee-Whiz Graph. Consider Figure 10; note the rise in ages of tetanus victims. Suppose someone thought the rise embarrassing and wanted to obscure the effect. All he need do is shorten the Y axis by changing its scale and extend its range a bit, as in Figure 11. Now the line seems not to ascend as much as before, even though the Y values are the same. And consider how much more chicanery is possible by throwing out the years 1952 through 1966, leaving only the data from recent years, which show us very little about the long-range, upward trend.

Or suppose the opposite effect were desired, i.e., to make the rise in ages seem alarmingly rapid. Simply lengthen the Y axis by adjusting its scale, put more tic marks along it, and chop it off at the bottom. Figure 12 shows the (somewhat misleading) results.

Either deception would be even worse if the Y axis were purposefully not labelled. Then the message in the data would be hopelessly obscure. The poor reader would be helpless. He cannot understand what isn't there, and his opinions become even more vulnerable to manipulation.

FIGURE 9

NORTH CAROLINA ESTIMATED POPULATION, JULY 1, 1979

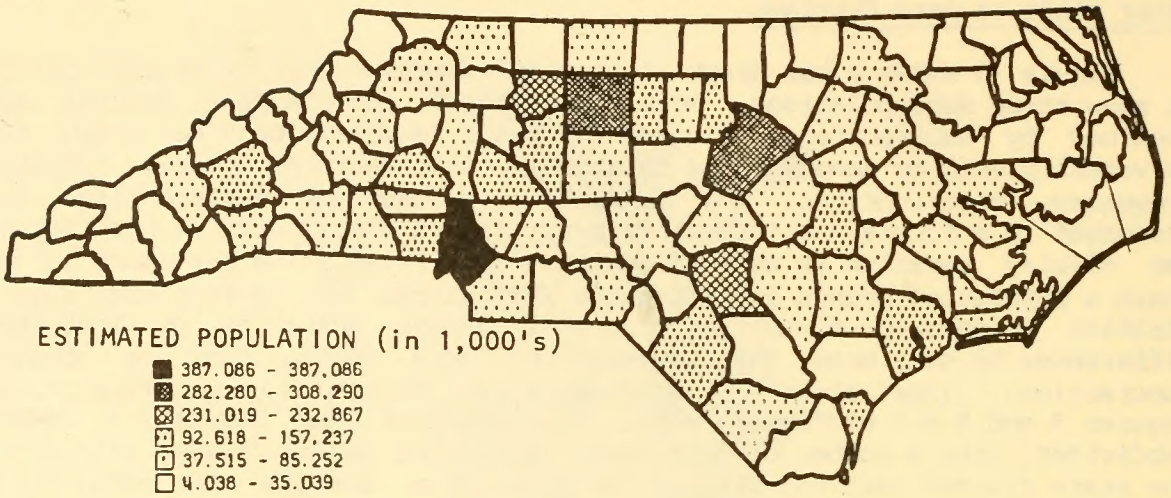


Figure 10

NON-NEONATAL TETANUS

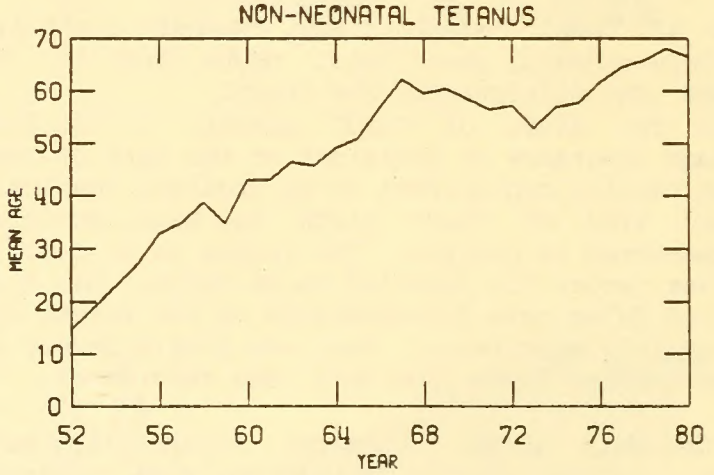


Figure 11

NON-NEONATAL TETANUS

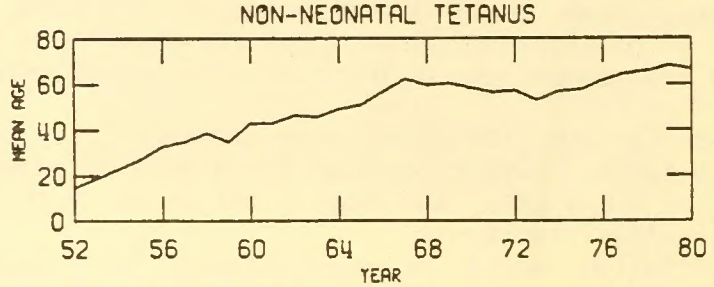


Figure 12
NON-NEONATAL TETANUS

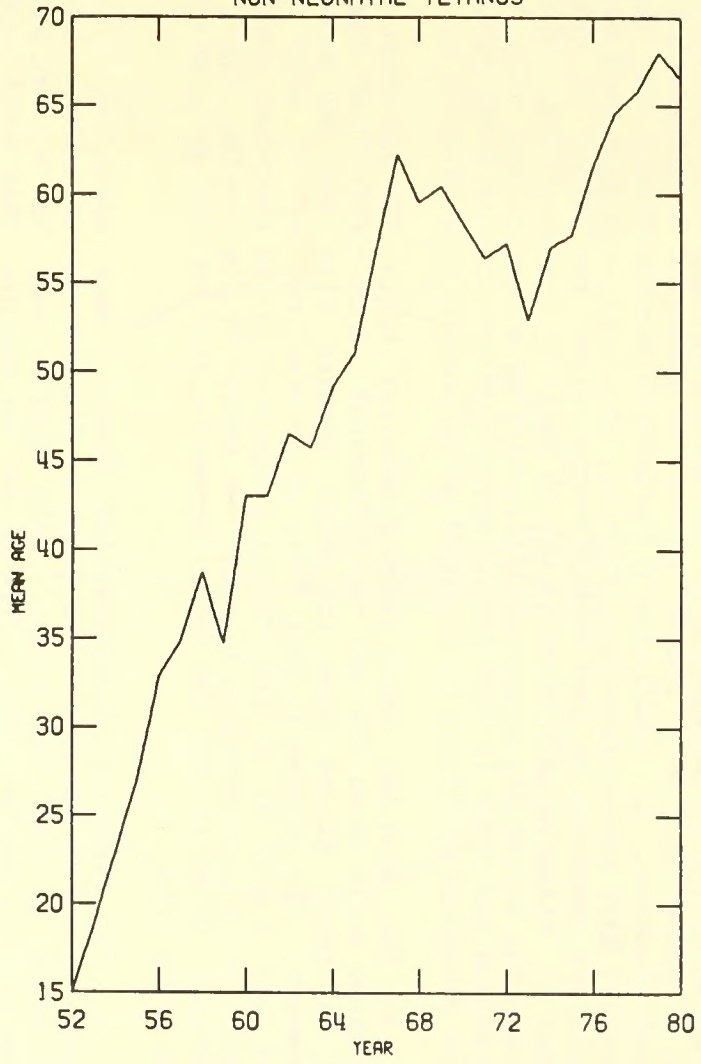
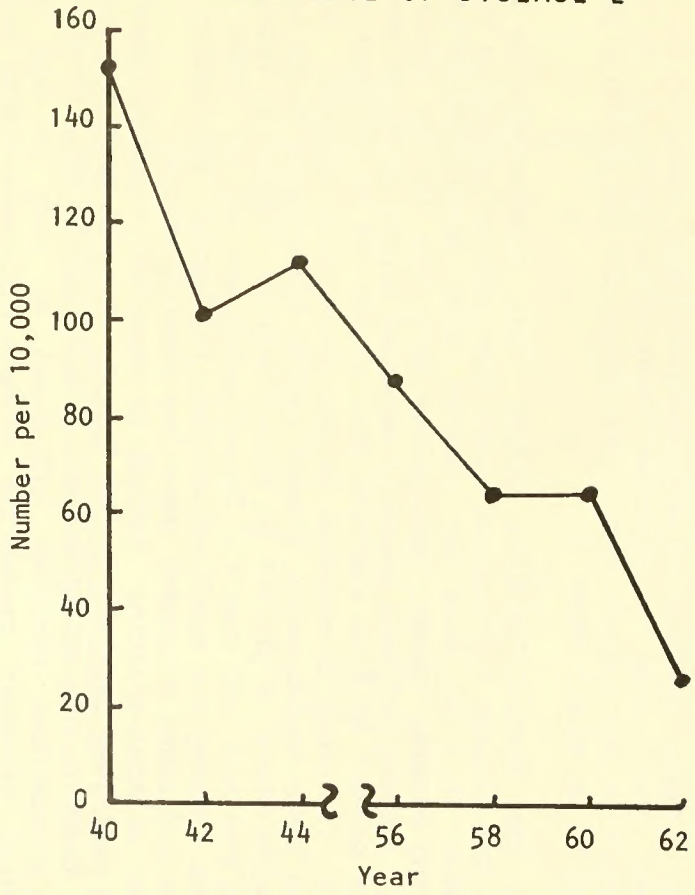


Figure 13
INCIDENCE OF DISEASE E



All graphs confront us with the problem of choosing the "right" scales for the axes. There exist no objective standards to guide us, so we have to rely on our intuition and subjectivity. The lesson of the gee-whiz graph is that we are supposed to be honest about it.

2. The Wishful-Thinking Graph. Our concern here is incomplete data, or rather, the blatant fabrication which some authors resort to in their attempts to deal with it. We have no quarrel with statistically valid procedures for estimating missing values. Incomplete data can be an embarrassment (if the missing numbers indicate a poorly designed survey or experiment), a vexation (if the gaps prevent us from learning what we set out to discover), and a budgetary disaster (if money is required to get the missing values). Some people, seeking to avoid these consequences, cannot resist the temptation to pretend missing data are really there, a pitfall of interpolation. Figure 13 illustrates the sort of make believe that can result. There really are not any data between 1944 and 1956, but fantasy has supplied what reality lacks. The author should have broken the line and added a comment about the missing values.

The wishful-thinking fantasy is especially serious when the X axis is time and the contrived data are in the future. A graph that ventures into the future is an example of extrapolation. Such graphs are, in effect, predicting the future, a hazardous exercise but legal so long as the assumptions--linear regression, constant percentage change, etc.--are stated and proper procedures are followed. Deception occurs when a graph goes off into the future and no assumptions are given. In some cases the assumption appears to have been that the future will be nice and rosy.

A related deception arises when the separate points of discrete data are connected as though they were continuous. (Roughly speaking, discrete data may be thought of as measurements that exist at only a few values of X; the X values are usually evenly spaced, as in Figure 1, and the Y values plot as separate points. Continuous data may be thought of as measurements that exist at all values of X; the X variable can assume every fractional value in its range, and the Y values plot as a "smooth" curve. We apologize to mathematicians for this casual treatment of continuity.) Ordinary, honest graphs of discrete data typically have marks--asterisks, circles, dots, etc.--at each data point; and the points are connected by line segments that give the graph an angular look. Figure 1 illustrates this valid way of handling discrete data. Trickery creeps in when the points are connected arbitrarily by a curved line that smooths out the angles and makes it seem that a continuous variable is being plotted. Figure 14 is a version of Figure 1, doctored up to produce this illusion of continuity.

3. The Much Ado about (Nearly) Nothing Chart. The pie chart in Figure 15 has three small areas that are misrepresented. Look carefully at countries A, B, and C; their slices have too much angular size for the percentages they represent. The problem is that there is no practical way to draw values like 0.1% and 0.06% on a pie chart; a single line has thickness greater than the angular width corresponding to 0.1%. So beware of pie charts. Small slices are sometimes drawn to give the false impression that certain parts are greater than they truly are.

4. The Cumulation/Dissimulation Graph. Look again at Figure 7. As a display of age-by-sex deaths the graph is dramatic and appealing. But it gives a misleading initial impression: male deaths seem to be, at all ages, more numerous than female deaths. For age 22, for example, most people would get the

Figure 14

"SMOOTHED-OUT" VERSION OF FIGURE 1

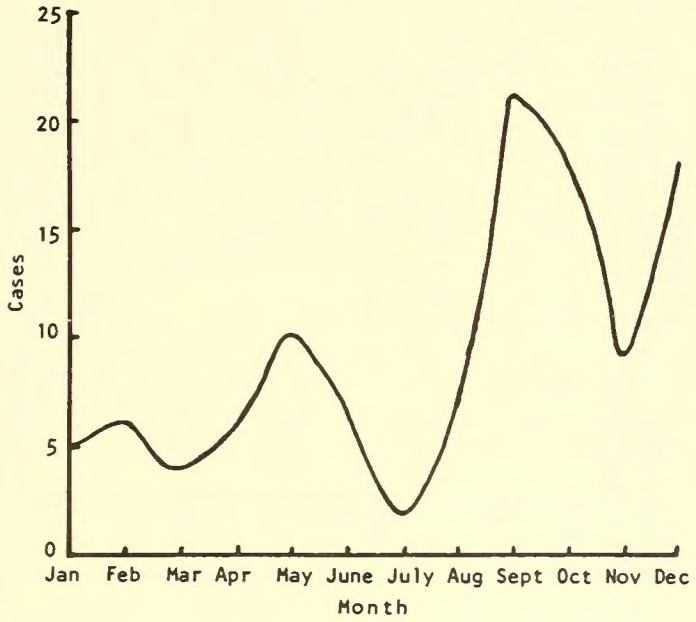


Figure 15

DISTRIBUTION BY COUNTRY OF BIRTH
OF IMMIGRANTS, 1905-1910

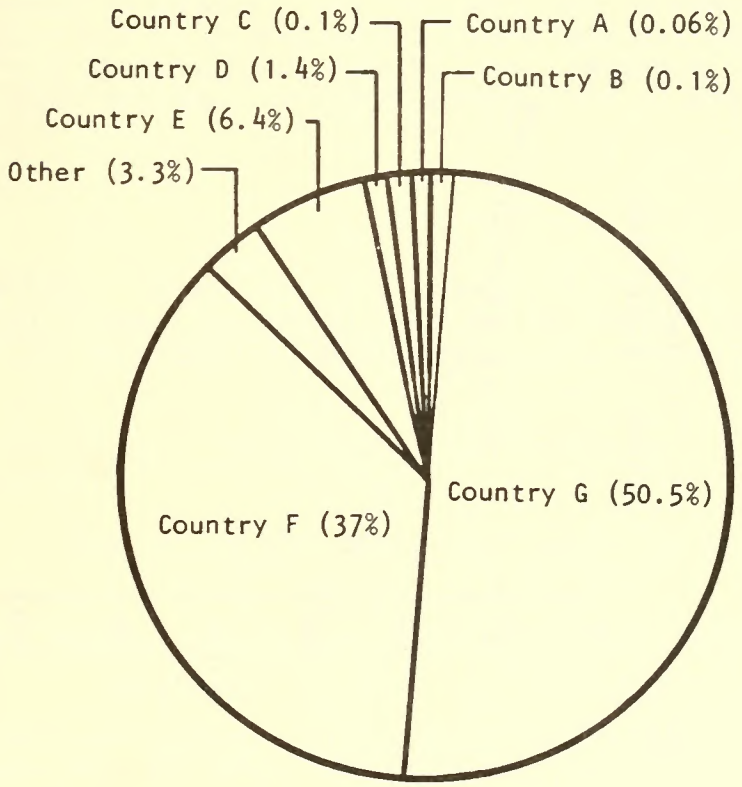


Figure 16
BIRTHS, 1955-1980

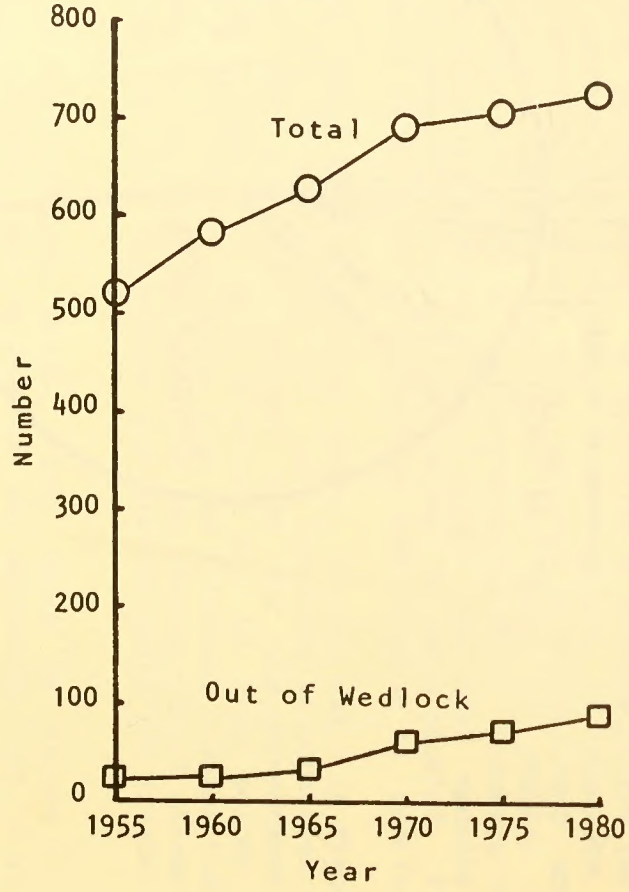
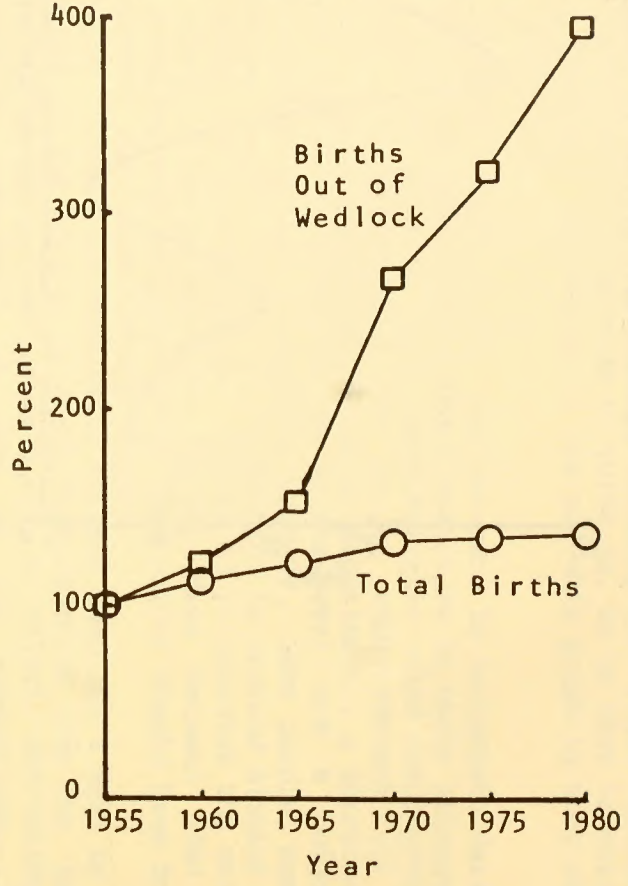


Figure 17
BIRTHS, 1955-1980, NUMBERS AS PERCENTAGES OF THE BASE YEAR 1955



immediate impression of about 230 female deaths (correct) and 380 male deaths (incorrect). Actually, there were only about 170 male deaths. The age distribution of male deaths is fitted above the female distribution, so the upper boundary of the "male zone" is the cumulation of male and female deaths. This feature is not obvious, and most people need to have it pointed out to them. The problem could be largely removed if the legends were annotated to read "Female, read directly", "Male, by subtraction", and "Total, read directly".

The same sort of thinking is required to interpret the cumulation bar chart in Figure 3. The legends should make it clear that the "Rural" and "Added" categories can be measured only by subtraction.

5. The One-Sided Coin Graph. The last warning we will sound is exemplified by Figures 16 and 17. Note that both graphs present the same data, first as births counted directly, then as births measured as percentages of the number of 1955 births, the base value. There is nothing wrong with either presentation. Taken together, the two graphs show clearly how total births and out-of-wedlock births increase.

But be careful when one of the graphs is purposefully not shown. Someone may be trying to trick you, especially if there is a controversial subject behind the data. One graph may appear to support one side of an argument, while the other graph may appear to support the other side. If the reader were shown both pictures he might realize the subject is not open-and-shut, and he might not form the "right" opinion.

Conclusions

Public Health is a world of data. All of us in that world should be skilled in presenting and interpreting numbers, but we must not practice the skill naively. Users of data have to be wary of misleading data displays, either due to accidental oversight or to intentional deceit. And producers of data who intend no deceit must conscientiously strive to present data in ways that give an honest impression.

References

Dyal, William W.; Eddins, Donald L.; and Peavy, J. Virgil. Descriptive Statistics. Tables, Graphs, and Charts. Center for Disease Control; Public Health Service; U.S. Department of Health, Education, and Welfare; Atlanta, Georgia 30333.

Ehrenberg, A. S. C. (1981). "The Problem of Numeracy". The American Statistician, 35:67-71.

Huff, Darrell (1959). How to Lie with Statistics. W. W. Norton and Company, Inc., New York.

Spear, Mary E. (1952). Charting Statistics. McGraw-Hill Book Company, Inc., New York.

State Center for Health Statistics. North Carolina Vital Statistics, Volume 1. North Carolina Department of Human Resources, P.O. Box 2091, Raleigh, N.C. 27602.

NORTH CAROLINA
Department of Human Resources
Division of Health Services
State Center for Health Statistics
P.O. Box 2091, Raleigh, N.C. 27602

Robert C. Overby, Jr.
Analyst Programmer
State Data Center, Budget Office
116 W. Jones Street
Raleigh, NC