GRAPHICAL APPROACH TO STATISTICS

by
C. J. Velz,
Prof. & Chm.
Dept. Pub. H'lth. Stat.
School of Pub. Health
UNIVERSITY OF MICHIGAN

Reprinted from
WATER & SEWAGE WORKS

Foreword

THE proper evaluation of data in sewage works operation and stream pollution studies cannot be accomplished by simple mathematical averages of a series of analytical results. "Statistical evaluation offers the only means of condensing raw data to concise form, permitting significant interpretation," according to the author.

These articles on the "Graphical Approach to Statistics" originally appeared in WATER & SEWAGE WORKS Magazine to show how the usual forbidding methods of statistics can be used by means of graphic techniques.

These articles may well improve future work in this field with particular reference to water and sewage analysis and bacteriological constituents.

Contents

Reprinted From and Copyrighted by WATER & SEWAGE WORKS Magazine,
a Scranton Gillette Publication

GRAPHICAL APPROACH TO STATISTICS

By C. J. VELZ

Professor and Chairman, Dept. of Public Health Statistics School of Public Health, University of Michigan, Ann Arbor, Mich.

I—Nature and Variability of Data

SCIENCE and engineering today are all but buried beneath the mass of their own data. In fact, the mere accumulation of more and more data is too frequently regarded as research. Statistical evaluation offers the only means of condensing raw data to concise form, permitting significant interpretation. Statistics provides the tools sharp enough to dig into the old accumulations, extricate and bring to light some of the buried "gold."

It is the purpose of this series to encourage the use of statistics by presenting a few simplified graphical techniques readily usable in Sanitary Engineering, i.e. procedures employing different types of probability papers, graphical tests for statistical significance of series of data and trends, and evaluation of laboratory results.

To illustrate the uses of statistics in Sanitary Engineering consider one problem, that of interpretation of suspended solids determinations at a sewage treatment works. We all know that suspended solids are subject to considerable variation from day to day and from hour to hour. How can we best define and summarize these variations and detect real changes in the load? When is a change significant beyond that expected by chance variation alone? How many individual determinations must we make in order to have a reliable average? When are individual determinations out of line with what is normally expected? How can we make reliable comparisons between influent and effluent?

For intelligent application of the graphical procedure, it is first necessary to gain fundamental concepts of the nature of variation of data and the laws of probability.

Definition

Statistics may be defined as the scientific collection and analysis of data, and the projection of estimates therefrom. Statistics is not an end in itself; rather it is an aid to judgment in arriving at valid conclusions, testing theories, measuring phenomena, discovering relationships, or projecting estimates under different conditions. The statistical method is an economical procedure for getting at the facts and determining their significance with confidence. In our enthu-

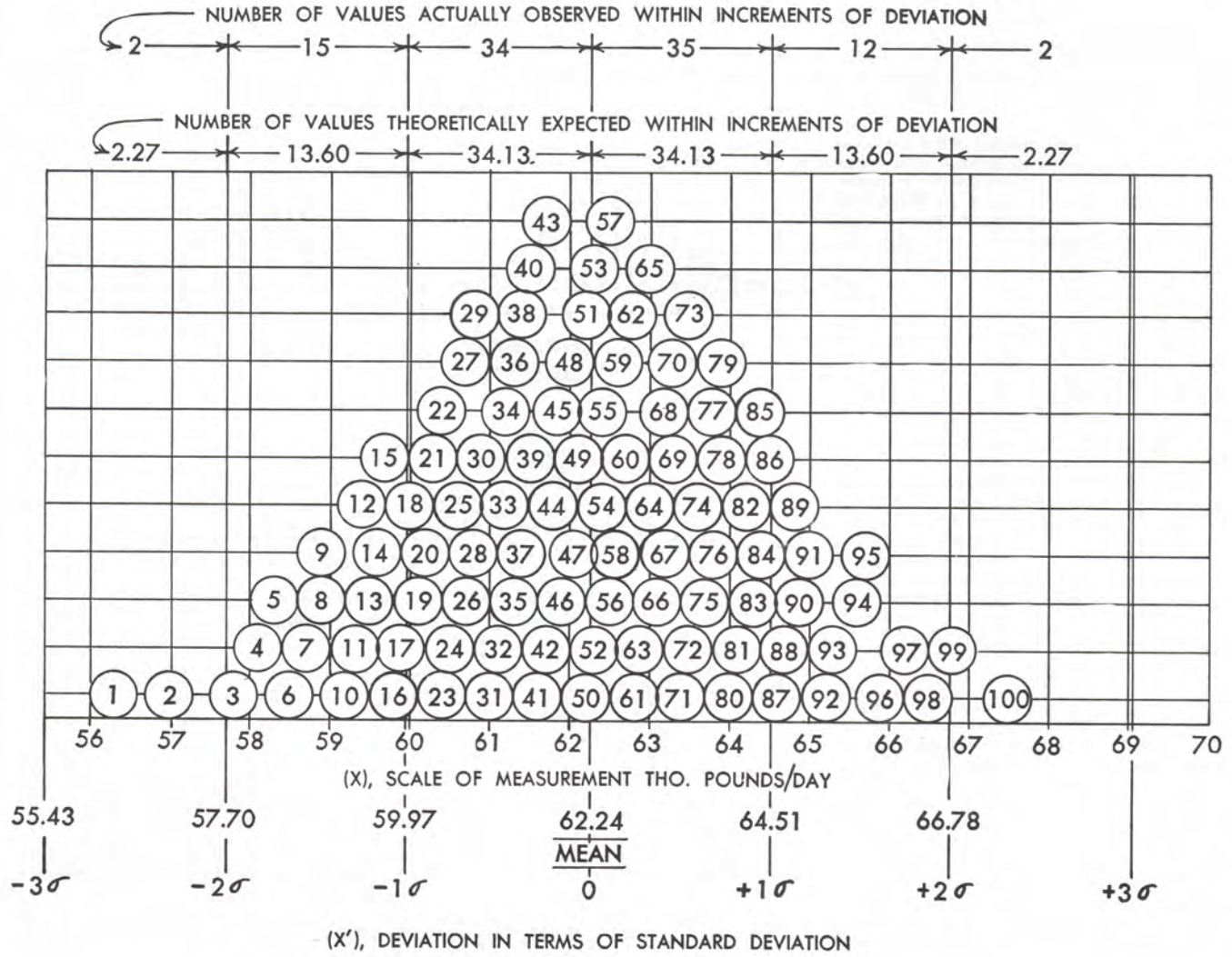


Fig. 1—Demonstration of the Normal Probability Curve
(Distribution of 100 suspended solids determinations; Mean 62.24 thd. pounds per day; Standard deviation 2.27 thd. pounds per day)

siasm over statistics it must be remembered that it is only a tool and is not a substitute for professional knowledge and experienced judgment. Sound judgments, however, cannot be made with confidence without the use of this tool.

Statistical Concepts

The whole concept of statistics is one of *variation*. If there were no variation in data there would be no need for statistical methods. This concept of variation is in conflict with most of the mathematics as it has been taught us. In mathematics we usually start with a preconceived theory which, if followed, automatically must lead to a single exact answer without any deviation. Such inflexible finality is not consistent with the facts of life as we *observe* them. Even with the most precise methods of measurement there is *always variation*. In statistical reasoning we start with observations and with full acceptance of their inevitable variation which leads not to a single answer but rather a range of answers.

For example, let us suppose we desire to know the weight of an object. A single careful measurement may be 10.1 milligrams, a second measurement made under the same circumstances may be 9.8, and similarly, repeated measurements will show other variations. With these variations in measurements the question arises, Which of these answers is correct? or, What is the exact true weight? The true weight remains unknown and at best can be only estimated. The best estimate is the average of the repeated measurements, in this case, say 10.0. But confidence in any estimate can be had only in defining not a single value but a *range* within which we believe the true weight lies.

From a statistical analysis of the variation of the measurements confidence and range can be related. We are confident of enclosing the true value 68 times in 100 within the range $10.0 \pm .1$ (9.9 to 10.1); or we are confident of enclosing the true value 997 times in 1,000 within the range $10.0 \pm .3$ (9.7 to 10.3). In the second instance our confidence is high; only 3 times out of 1,000 would we be wrong in believing that the true value was within this wider range. But we have gained this added confidence only at the expense of precision; we are less definite as to the exact true weight in the sense of a single value.

In fact, under this new concept we can never determine with finality the exact true weight, because as we narrow the range working toward a single value we lose confidence in enclosing it, with total loss of confidence if an attempt is made at expression as a single value. Thus in expressing measurements we always state a *range;* the mean plus or minus some increment of standard deviation, the increment dependent upon the degree of confidence desired.

The Nature of Variation

If a series of measurements, say weight, made on the same object are sorted in order of magnitude, it will be found that the results vary in a quite systematic manner. Many of the results will cluster near the midvalue and a few will be more or less evenly distributed with greater spread above and below the central majority. Another type of quantitative data consisting of a *sample* from a large population, universe or supply such as a single weight measurement of a number of such objects selected at random, will likewise portray this tendency toward central cluster and symmetrical spread.

If a large number of such measurements are made and plotted on the scale of measurements, each individually represented by a circle, a symmetrical bell-shaped pile develops as shown in Fig. 1. The center of each circle is vertically above the magnitude of its measurement on the (X) scale; the number in each circle is the rank of each measurement arranged in ascending order of magnitude. The measurements cluster so closely about the mean that it is necessary to pile up the circles in order to locate them opposite their position on the (X) scale, thus forming the bell-shaped curve or distribution. This frequency curve of distribution of measurements, referred to as the *Normal Probability Curve,* has definite characteristics with respect to the proportion of the measurements found in equal increments of scale as indicated by the numbers across the top of the pile. Those characteristics will be defined more precisely later.

Suffice to observe that the distribution is nicely symmetrical about a *centering value;* with the *deviations* above and below this value spreading out in a very orderly fashion. Other random samples taken from the same universe under the same conditions will reproduce quite similar distributions both as to centering position on the scale and as to deviation or spread. Measurements of different phenomena will develop similar bell-shaped distributions but it will be observed that each has its own centering value and that some vary in the degree of spread or deviation about their central values as illustrated in Fig. 2.

Centering Values

Three centering values are commonly employed: the *Median,* the *Mode* and the *Mean.* The median value is the value exceeded by ½ of the values and of which ½ of the values fall short. It represents the series, if at all, by virtue of its position of dividing the series in half. The mode is the value in the series most frequently repeated and around which other values cluster most densely. It is the high point in the distribution irrespective of the number of terms above and below it. The mean is a simple arithmetic average of the series of data represented by the *sum of the individual measurements divided by the number of observations.* Where the distribution is *normal* (symmetrical), the median, the mode and the mean are identical and coincide. The mean is the most useful centering value and is most frequently employed.

However, an average by itself is a dangerous value. Too frequently series of data are summarized merely by an "average" (the mean) without designating the number of measurements or the nature of variation or spread in the data. For example, consider the following three series of data, X, Y, and Z, which have the same average:

(1) X comprised of 2 values, 1 and 99; average = 50.
(2) Y comprised of 19 values ranging from 10 to 90; average = 50.
(3) Z comprised of 19 values ranging from 47 to 53; average = 50.

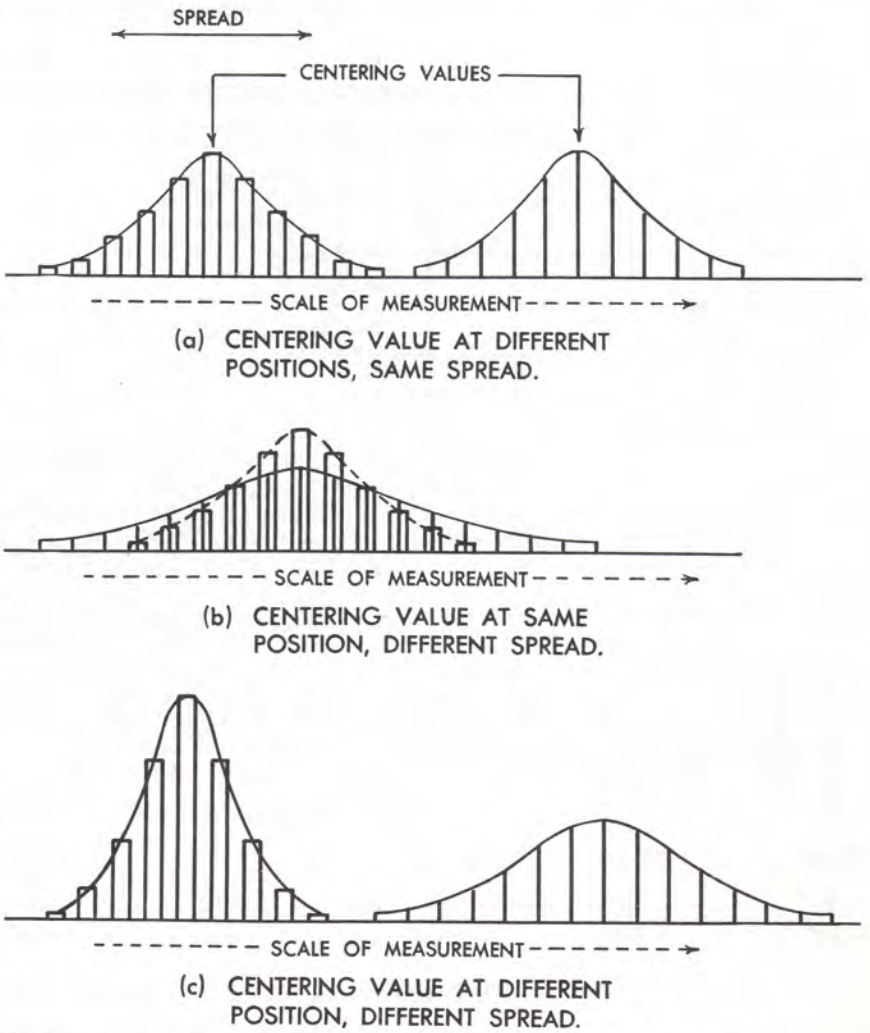


Fig. 2—Bell-shaped Distributions Show Centering Tendencies but Vary in the Degree of Spread or Deviation

While all produce the same average (50), it is not difficult to see that the three averages come from quite different data. In the first instance, two measurements hardly give meaning to an average, particularly when they have a great spread; in the second and third instances, while both have the same number of measurements (19) and the same average (50), the second has a wide spread ranging from 10 to 90, while in the third the data are compact, with a spread ranging from 47 to 53.

Deviation from the Central Value

In addition to the mean and the number of measurements it is necessary to know something of the nature of deviations of measurements about the central value.

The Standard Deviation, σ (Sigma), is the fundamental measure of variation which constitutes the major "building block" in all statistical procedures. It is the root mean square deviation obtained by squaring each individual deviation from the mean, summing, dividing by the number of measurements and extracting the square root of this mean square.

$$\sigma = \sqrt{\frac{\Sigma(X - \overline{X})^2}{n}} \qquad \text{Eq. 1}$$

(There are various other transformations of this equation which are more useful if computations are made by machine; also a simple graphical method of obtaining the standard deviation will be given in the next article.)

The standard deviation is important because, through it, the fundamental characteristics of the normal probability curve are defined and precision can be related to probability or a confidence level.

Probability

As a basis for understanding the graphical applications of statistical procedures, which will be presented in subsequent articles, and the significance of the standard deviation as a measure of variation, it will be necessary to gain a concept of *Probability.*

There are two concepts of probability, the one based upon what *can* occur and the other based upon what *has* occurred. If an event can occur or has occurred in (a) ways and can fail or has failed in (b) ways, then the probability of a success is $a/(a + b)$, which is a ratio of the number of ways in which an event can occur or has occurred to the total number of ways of occurrence, each being equally likely to occur. Probability is expressed as a ratio and ranges from *zero* to *unity;* zero designating impossibility; unity designating certainty.

The first concept, based upon what *can* occur, is referred to as *a priori* probability. In tossing a coin there are two ways in which an event can occur, either a head or a tail. The probability of a specified event, head, is therefore ½ or 0.5. There are six ways in which a die can fall. The probability of the event, an ace, is therefore 1/6. These are probabilities before a trial, based purely upon what *can* occur.

The second concept, based upon what *has* occurred, is referred to as *a posteriori* probability, probability after the trial. A coin is tossed 1,000 times and the event, a head, has been observed to occur 506 times. The a posteriori probability of a head, based upon what has occurred, is 506/1,000 or 0.506, as compared with a priori probability based upon what can occur of 0.5.

The a priori or true probability of a universe can exist but it cannot be measured. Repeated random samples drawn from a universe show results varying about the true value, results from the small samples showing greater variation than those from the larger samples. Or as the number of trials approaches infinity the a posteriori probability as determined from the observations approaches the true a priori probability.

Numerical Expressions of Probability

1. The sum of the probabilities of occurrence of a complete and mutually exclusive set of events is unity. Impossibility has a probability of exactly zero. The entire range of probabilities extends from impossibility to certainty or numerically from 0 to 1.

2. If the probability of success is p and of failure q then $p + q = 1$; and $1 - p = q$; and $1 - q = p$.

3. The probability of *either* one *or* another of mutually exclusive events occurring is the sum of their individual probabilities. Probability of drawing a spade *or* a heart: $\frac{13}{52} + \frac{13}{52} = 0.5$; of drawing a spade, a heart or a club: $\frac{13}{52} + \frac{13}{52} + \frac{13}{52} = 0.75$.

4. The probability of occurrence of a compound event composed of two or more independent events is the product of the individual probabilities. What is the probability of *two* heads in *two* throws of a coin? (One head on the first throw and one head on the second throw.)

Probability of head on 1st throw = 0.5
Probability of head on 2nd throw = 0.5
Probability of two heads on two consecutive throws $0.5 \times 0.5 = 0.25$

Relation of the Standard Deviation to the Normal Probability Curve

The normal probability curve is a standardized form of distribution where the scale of measurement is in terms of the standard deviation measured from the center of the distribution as the origin. For example, referring to Fig. 1, the (X) scale is weight in thousand pounds per day; and the corresponding (X′) scale is in terms of the standard deviation. The zero of the X′ scale is opposite the center of the distribution, the mean 62.24 thousand pounds per day. Unit distance on the X′ scale is equivalent to a standard deviation, 2.27 thousand pounds per day, measured along the X scale but labeled $+1\sigma$, $+2\sigma$, $+3\sigma$ and -1σ, -2σ, -3σ, on the X′ scale. To convert a series of scaled measurements (X) to the standard probability scale X′ it is necessary to know the mean $(\overline{X})$ and the standard deviation (σ). The mean $(\overline{X})$ is the simple arithmetic average and the standard deviation can be computed from the data or, as will be illustrated later, readily obtained graphically.

The significance of the standard deviation (σ) as a measure of variation is apparent from the following analysis of the distribution of the 100 measurements on Fig. 1. Extending two vertical lines at $+1\sigma$ and -1σ on either side of the mean, we count 69 of the total 100 measurements falling within this range, and 31 fall without, 14 above and 17 below. Probability of measurements falling within this range based upon what has occurred is 69/100 or 69 per cent; and of falling without, 31/100 or 31 per cent. Similarly, we count 96 measurements within the range of the mean $\pm 2\sigma$, and 4 measurements without, 2 above and 2 below. The probability of measurements falling within is 96/100 or 96 per cent, and of falling without, 4/100 or 4 per cent.

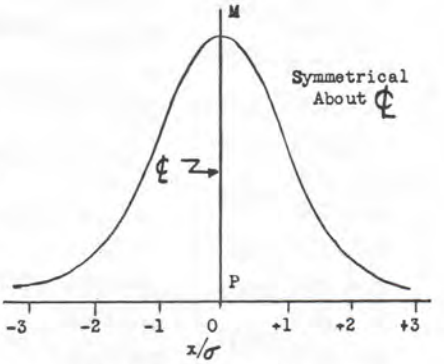
Characteristics of the Normal Probability Curve

These results are based upon a sample of 100 measurements. As the number of measurements is increased the a posteriori distribution approaches the a priori distribution defined by the Normal Probability Curve. The a priori probability of measurements falling within any range are given by the area under the curve between the limits of the range defined in terms of standard deviations from the mean. These areas have been computed for any increment of deviation and are available in tables.

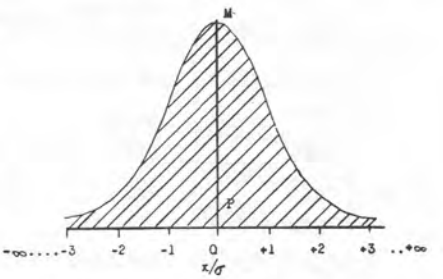
One useful form is the Integration Graph shown in Fig. 3. The areas are given between the central value (the mean) and x/σ (deviation in terms of standard deviation) measured only in one direction from the mean.

The Probability Curve has the following six characteristics:

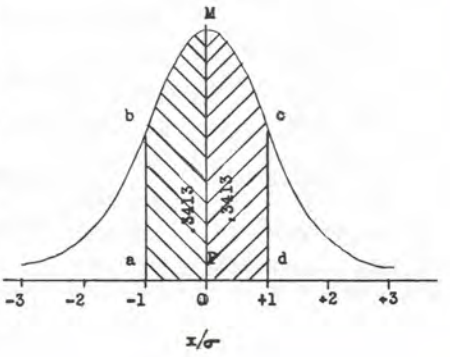
1. All events are distributed symmetrically about the "mean" or the "most probable" M.P. value.



2. The Probability Integral for the complete distribution of events is the area between $\pm \infty$ which is unity. Pb of all events = 1.0.



3. The Probability Integral for all events between the limits of $\pm 1\sigma$ (Dev/σ = 1) from the M.P. value is the area (abcd) which from the Integration Graph is $2 \times 0.3413 = 0.6826$. Thus 68.26% of all events lie within a range of $\pm 1\sigma$ from the M.P. value and 31.74% of the events lie without this range.



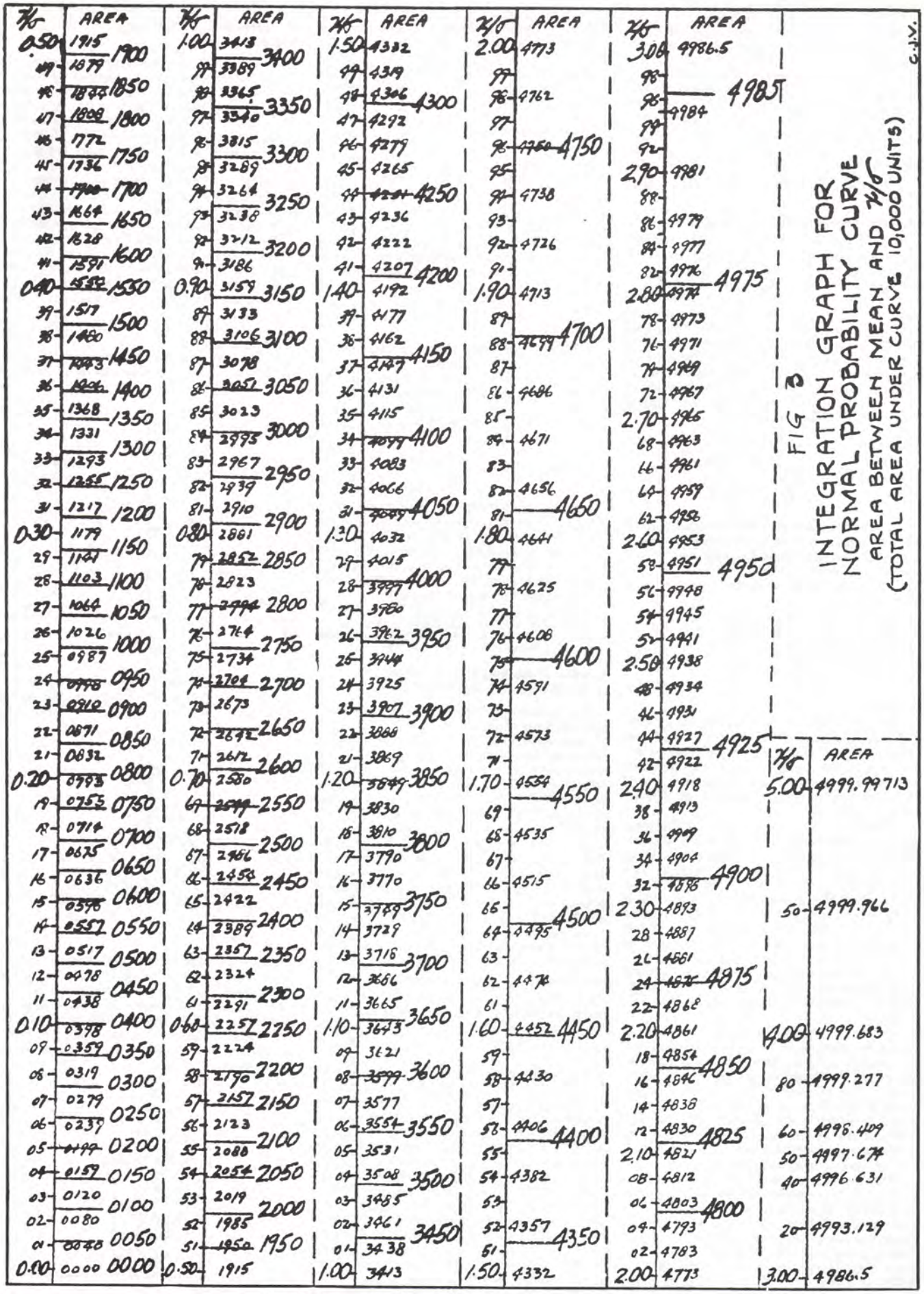
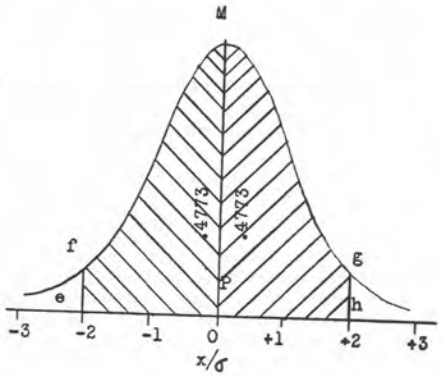
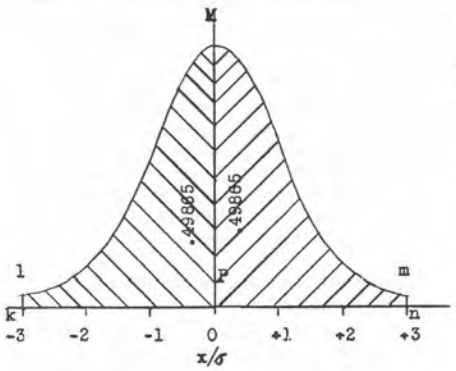


Fig. 3—Integration Graph for Normal Probability Curve
(Area between mean and ×/σ. Total area under curve 10,000 units)

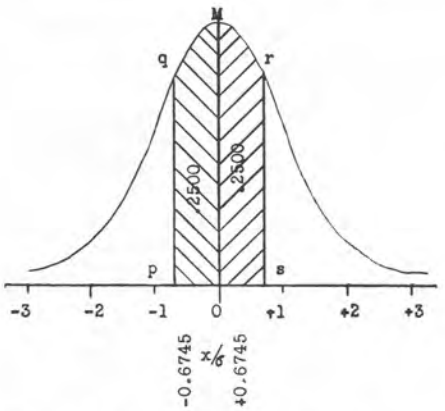
4. The Probability Integral for all events between the limits of $\pm 2\sigma$ (Dev/$\sigma = 2$) from the M.P. value is the area (efgh) which from the integration graph is $2 \times 0.4773 = 0.9546$. 95.46% of all events lie within a range of $\pm 2\sigma$ from the M.P. and 4.54% of events lie without this range.



5. The Probability Integral for all events between the limits of $\pm 3\sigma$ (Dev/$\sigma = 3$) from the M.P. is the area (klmn) which from the integration graph is $2 \times 0.49865 = 0.9973$. 99.73% of all events lie within the range of $\pm 3\sigma$ from the M.P. and 0.27% of the events lie without this range.



6. The Probability Integral for all events between the limits of $\pm 0.6745\sigma$ (Dev/$\sigma = .6745$) from the M.P. is the area (pqrs) which from the integration graph is $2 \times 0.2500 = 0.5000$. 50% of all events lie within the range of $\pm 0.6745\sigma$ from the M.P. and 50% of events lie without this range. This is sometimes referred to as the range of Probable Error (P.E.).



Uses of Probability Integration Graph

Let us suppose we wish to define the quality of effluent of a sedimentation tank in terms of suspended solids. The mean of 30 determinations is 102.0 ppm, but values vary about the mean with a standard deviation from the mean (σ) of 5.5 ppm. From what has been said heretofore, we must define quality of the effluent by a range and we wish to know the confidence of enclosing the quality in that range.

For example, suppose we make the statement that the quality of effluent is within the range of 102.0 ± 10 ppm.; what is our confidence that this range encloses the quality? or, what risk do we take in having the quality measurement fall outside this range? This we determine from the integration graph:

Deviation from the mean = × = 10 ppm;
Deviation from the mean in terms of standard deviation, $x/\sigma = \frac{10}{5.5} = 1.82$.

Opposite x/σ of 1.82 on the Integration Graph, Fig. 3, read area 46.56 per cent or probability for the range, Mean ± 1.82σ, is 2 × 46.56 per cent or 93.12 per cent.

We are about 93 per cent confident of enclosing the quality of the effluent within the range of 102.0 ± 10 ppm. In accepting this range as defining the quality we run the risk of being wrong about 7 times in 100, as 7 times in 100, deviations greater than ± 10 ppm are expected.

Again suppose we do not wish to be wrong in our range of quality more often than once in 100, then in 99 times in 100 we must be right. This requires an area in one direction from the mean of 99/2 or 49.5 per cent. Opposite area of 49.5 per cent on the Integration graph, Fig. 3, read $\times/\sigma$ of 2.567. The range then is defined by the mean $\pm 2.567\sigma$ or 102.0 ± 2.567 × 5.5, 102.0 ± 14.1 ppm. In other words, in saying that the effluent quality is 102.0 ± 14.1 we are 99 per cent confident of being right and only once in 100 would we expect to be wrong by having quality fall outside of this range.

If we are willing to accept even chance of being right or wrong, the range can be narrowed. This requires an area in one direction from the mean of 50/2 or 25 per cent. Opposite area of 25 per cent on the Integration Graph, Fig. 3, read $\times/\sigma$ of 0.6745. The range then is the mean ± $.6745\sigma$ or 102.0 ± .6745 × 5.5 ppm = 102.0 ± 3.7. Now we have a narrower range but we do not have great confidence in it because half of the time quality of the effluent is expected to fall outside the limits of this narrower range.

In the next article we shall transform the bell-shaped probability curve and its integration graph into a straight-line on probability paper so that answers to questions such as these as well as others can be directly read at a glance.

II—Normal Probability Paper

This section will deal with the construction and use of normal probability paper. The bell-shaped probability curve developed in the previous article can be reduced to a straight line in the form of probability paper which facilitates the graphical solutions of many statistical problems.

The first step in the construction of normal probability paper is the transformation of the bell-shaped distribution curve to the probability summation or the ogee curve. The bell-shaped curve and its integration graph, Fig. 3, give the areas of the central portion about the mean working outward toward the tail of the curve. The probability summation or ogee curve starts in the extreme left tail at minus infinity and sums the area by successive intervals toward the right. These two forms are shown in Fig. 4 (A), the bell-shaped distribution curve, and (B) the summation or ogee curve, both referred to an X′-axis scale in deviations from the mean expressed in terms of standard deviation.

The ogee curve can be constructed readily from the increments of area obtained from the integration graph, Fig. 3. The percentage of the area up to any deviation below the mean is simply obtained by sub-

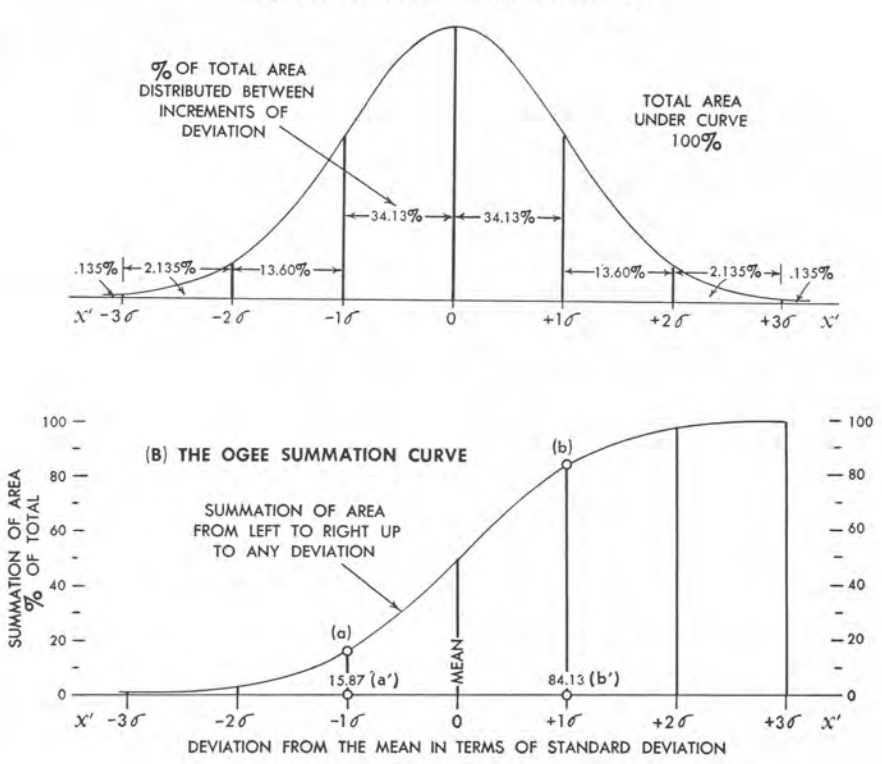


Fig. 4—The Normal Probability Curve

tracting from 50 per cent, the area obtained from the integration graph for that deviation; and for deviations above the mean, by adding 50 per cent to the areas of the integration graph, Fig. 3. For example, the summation of area up to a deviation of -1σ below the mean is (50 — 34.13) or 15.87 per cent, while the summation of area up to $+1\sigma$ above the mean is (50 + 34.13) or 84.13 per cent.

These two points are located on Fig. 4 (B) at (a) and (b) and similarly any number of other transformations can be made, forming a complete smooth ogee curve.

The next step is to reduce the ogee probability summation curve to a straight line. This is simply done by projecting the points on the ogee curve vertically to the linear X′-axis scale and writing opposite each its corresponding percentage. For example, point (a) is projected to (a′) and writing its probability summation 15.87 per cent; (b) projected to (b′) and writing 84.13 per cent; etc. Figure 5 (A) is such a scale enlarged and Fig. 5 (B) is a scale of a size convenient for transfer to an 8½ x 11 in. sheet. These scales constitute the X-axis of *normal probability paper.*

Since the probability summation is the cumulative area from left to right, it represents probability *equal to or less than.* This is a useful expression as it affords an opportunity to determine readily the probability of not exceeding a certain magnitude of measurement.

For example, arranging the series of data of suspended solids determination of sewage (Table 1) in ascending order of magnitude automatically places the data in the order of their probability *equal to or less than.* Thus, by the simple technique of arranging data in order of magnitude we place them in the order of their position on the probability summation scale of normal probability paper.

If the distribution is truly normal, these values so arranged form a perfect straight line on the paper. Obviously, it is easier to deal with a straight line than the complicated bell-shaped curve. There remains, however, the problem of determining the plotting position of a number of measurements on the probability summation scale.

The Plotting Position

Let us refer for a moment to the definition of probability given in the first article, namely, the ratio of the number of occurrences divided by the total number of trials. Applying this to the series of 20 determinations of suspended solids of sewage shown in Table 1 leads to the following:

CASE I. The probability of a value equal to or less than the smallest value, the first, would be 1/20 or 5 per cent; two values were observed equal to or less than 103.4 thousand pounds, probability 2/20 or 10 per cent; and finally 20 values are equal to or less than the largest value or 20/20 or 100 per cent. Immediately we face the dilemma, if we plot the first point at 5 per cent on the probability summation we will not be able to plot the last point because it is positioned at 100 per cent which is located at plus infinity, completely beyond our scale. This is equivalent to saying that the largest value which was actually observed to occur once in 20 would occur only once in an infinite number of observations.

CASE II. Similarly, if we reverse the scale and start at the large end and number the values in descending order of magnitude and express probability as equal to or greater than, the largest value would be plotted at 1/20 or 5 per cent, but now we would be unable to plot the smallest value as its probability would be 20/20 or 100 per cent which would be located at minus infinity, completely off the scale.

CASE III. Hazen[1] suggested what appeared to be solution; plot the first value at the midpoint of equal increments on the scale of 100 per cent. For example, the increment for the 20 points would be 100/20 or 5 per cent. The first point being plotted at ½ of 5 or 2.5 per cent; the second point at 2.5 + 5 per cent or 7.5 per cent; etc., successively to the 20th point at 97.5 per cent. This permits plotting all of the values on the scale, and at first thought appears to solve the problem; but this does violence to the probability of the extreme values, the first and last points. Placing the smallest value of 20 values in a position on the probability scale at 2.5 per cent is ascribing to it a probability as though it were the smallest value of a series of 40 measurements, 100/2.5. In like manner, placing the largest value of 20 at 97.5 per cent is to ascribe to it a position as though it were the largest value of a series of 40 measurements. Obviously, such positions of the probability of the end points cannot be accepted.

CASE IV. Theoretically, the plotting position for the mean of any series regardless of number of values is at 50 per cent and ½ of the values should be above and ½ below. If the mean is considered as part of the series the number of plotting positions becomes n + 1. On this basis the mean plotting positions of the observed values are obtained from the relation $\frac{m}{n+1}$ in which m is the serial number of the measurements arranged in ascending order of magnitude and n is the total number of measured values to be plotted. Thus, the plotting position for the first value of a series of 20 values would be 1/21 or 4.77 per cent; the 2nd, 2/21 or 9.53 per cent, etc.; and the 20th value, 20/21 or 95.23 per cent.

Under these conditions it will be noted that the first and last values in the series of 20 will be plotted in positions as though they are the smallest and largest values in a series of 21 values (100/4.77), with 10 values above 50 per cent and 10 values below 50 per cent. These plotting positions are more consistent with theory and with the facts as observed, and Case IV is therefore recommended.

[1]Hazen Allen. *Flood Flows,* John Wiley & Sons. 1930.

Plotting on Normal Probability Paper

In summary, the procedure for plotting observed or experimental data on normal probability paper is as follows:

1. Arrange the data in order of ascending magnitude.

2. Assign a serial number "m" to each of the n values 1, 2, 3.....n.

3. Compute the plotting position of each serial value, giving the probability equal to or less than for each value by the expression $\frac{m}{n+1}$ expressing the ratio as a percentage.

4. The scale by which the observed or experimental data were measured is then laid off on the Y-axis and the plotting positions are the probability scale on the X-axis. The points are plotted, using these (X, Y) coordinates.

Interpretation:

1. If a straight line develops in the plotting, it indicates that the data have a normal distribution; that is, in accordance with the theory of probability we expect results distributed in this manner.

2. If a straight line does not develop in the plotting, we suspect that something is changed in the conditions affecting the observed measurements beyond what is normally expected. This might mean carelessness in measurements or it may mean that we had not measured the same characteristics under the same conditions or that we had been measuring at the same time two or more distinctly different things. These very useful interpretations are immediately visually apparent in the graphical presentation on probability paper.

Graphical Determination of the Mean and the Standard Deviation

Following the above procedure, the 20 determinations of suspended solids of sewage shown in Table 1 are plotted on normal probability paper, Fig. 6. Inasmuch as the distribution forms a straight line (is normal), it is readily possible to determine graphically (1) the mean of the series and (2) the standard deviation from the mean, measuring the degree of variation among the data.

The mean is at once determined where

Table 1

Suspended Solids In Daily Raw Sewage Flow

(Data courtesy Geo. Fynn and Geo. E. Symons, Buffalo Sewer Authority)

Suspended Solids 1000 lb./Day arranged in order of magnitude	Rank Serial No. (m)	Plotting position, Probability—Equal to or less than (m/(n+1) in %
(1)	(2)	(3)
72.6	1	4.8
103.4	2	9.5
112.8	3	14.3
117.1	4	19.0
120.7	5	23.8
127.9	6	28.6
128.9	7	33.3
131.3	8	38.1
135.4	9	42.9
137.5	10	47.6
143.9	11	52.4
148.4	12	57.1
148.8	13	61.9
153.9	14	66.7
155.2	15	71.4
170.9	16	76.2
176.0	17	81.0
185.0	18	85.7
188.6	19	90.5
194.2	20 = n	95.2

the straight line of the distribution intersects the vertical line extending from 50 per cent on the X-axis at (a). The magnitude of the mean is read directly on the Y-axis as 143 thousand pounds per day.

The standard deviation is the slope ($\Delta y/\Delta x'$) of the distribution line on probability paper. If there were no variation in the individual measurements, all values would be exactly alike and would plot exactly as a horizontal straight line. The angle from the horizontal position, therefore, is a measure of the degree of variability of the data; the steeper the slope the greater the variability.

Since the probability summation scale was constructed linearly in terms of standard deviation from the mean, the slope of the distribution line measured on the standard deviation scale at once provides the standard deviation. Extending vertically from $+1\sigma$, one standard deviation from the mean, Δy, can be read directly from the graph on the Y scale as 30 thousand pounds per day. (The slope (σ) can be determined from any corresponding $\Delta y/\Delta x'$ along the line.)

In addition to visually summarizing our data, graphically supplying the mean and the standard deviation from the mean, plotting on normal probability paper affords an opportunity to extend our series beyond the number of observations and determine the expected magnitude of a value at any probability equal to or less than. For example, referring to Fig. 6, if we wish to know the expected magnitude of a value that not more than 1 per cent of the values are equal to or greater than, that is 99 per cent equal to or less than, we extend vertically from 99 per cent on the X-axis to intersect the distribution line at (b) reading on the Y-axis 212 thousand pounds per day.

We also can make interpretations concerning confidence and precision in any range of deviation about the mean. For example, we are 68 per cent confident of enclosing the quality within the range of 113 to 173 thousand pounds per day, located directly on the distribution line at the points (c) and (d), vertical extensions from 16 and 84 per cent, that is 34 per cent on either side of the mean. Similarly, we are 90 per cent confident of enclosing the quality within the range of 93 to 193 thousand pounds per day, located by (e) and (f), vertical extensions from 5 and 95 per cent on the X-axis, 45 per cent on either side of the mean.

Thus from the plot on probability paper we can read directly the probability of any range of deviation or the probability of not exceeding any magnitude of measurement.

Plotting a Large Series

The example of Fig. 6 deals with a small number of measurements, and no difficulty is encountered in plotting each individual measurement. In the event the series of measurements comprises a large number of values it becomes confusing to plot each individual value, and it is preferable to *group* the data. For example, suppose we have taken during a year 500 measurements of hardness of a softened water supply and we wish to summarize the data and define the variation by plotting on probability paper. The scale of measurement in ppm. has been divided into 5 ppm. intervals and the hardness results have been sorted into each group as shown in Table 2.

Summing the number of values found in each group gives the number less than the

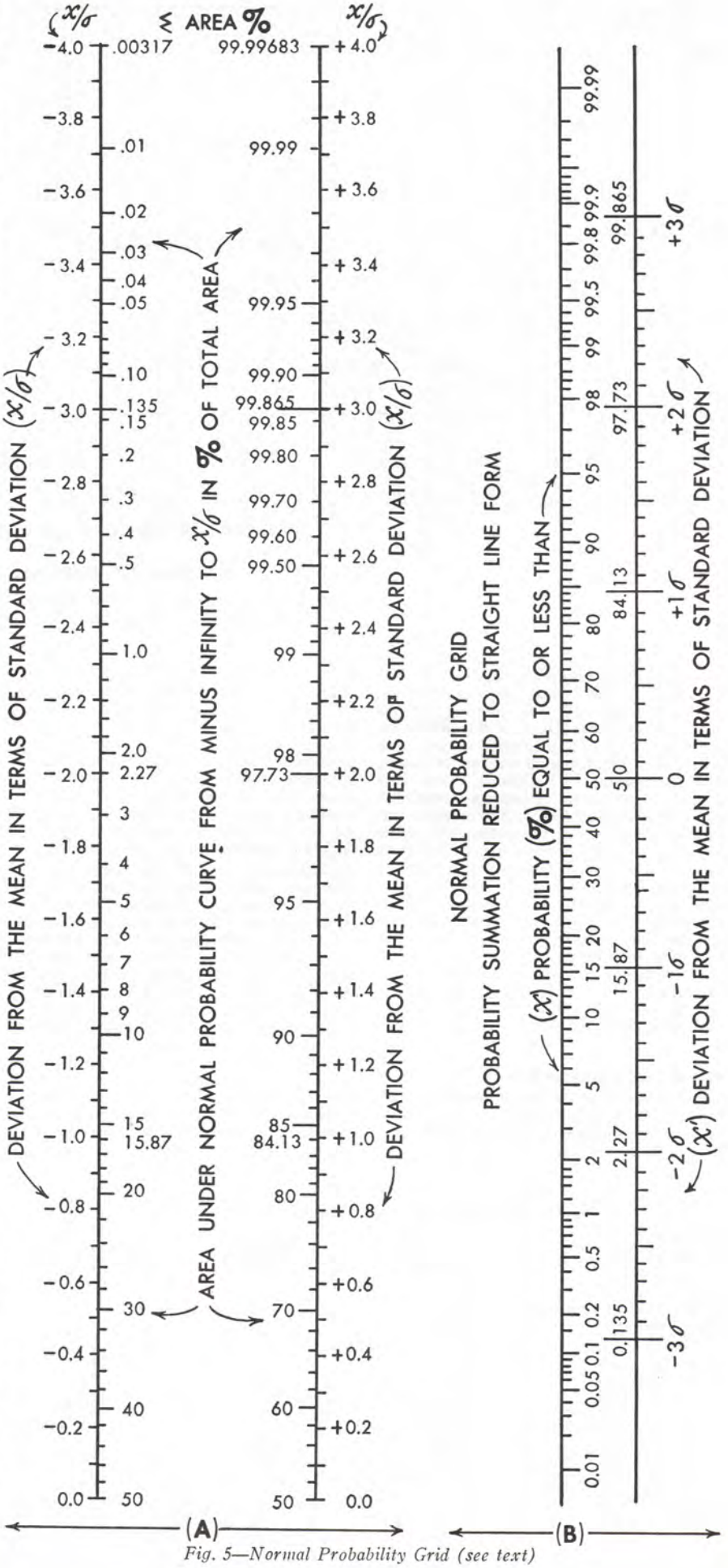


Fig. 5—Normal Probability Grid (see text)

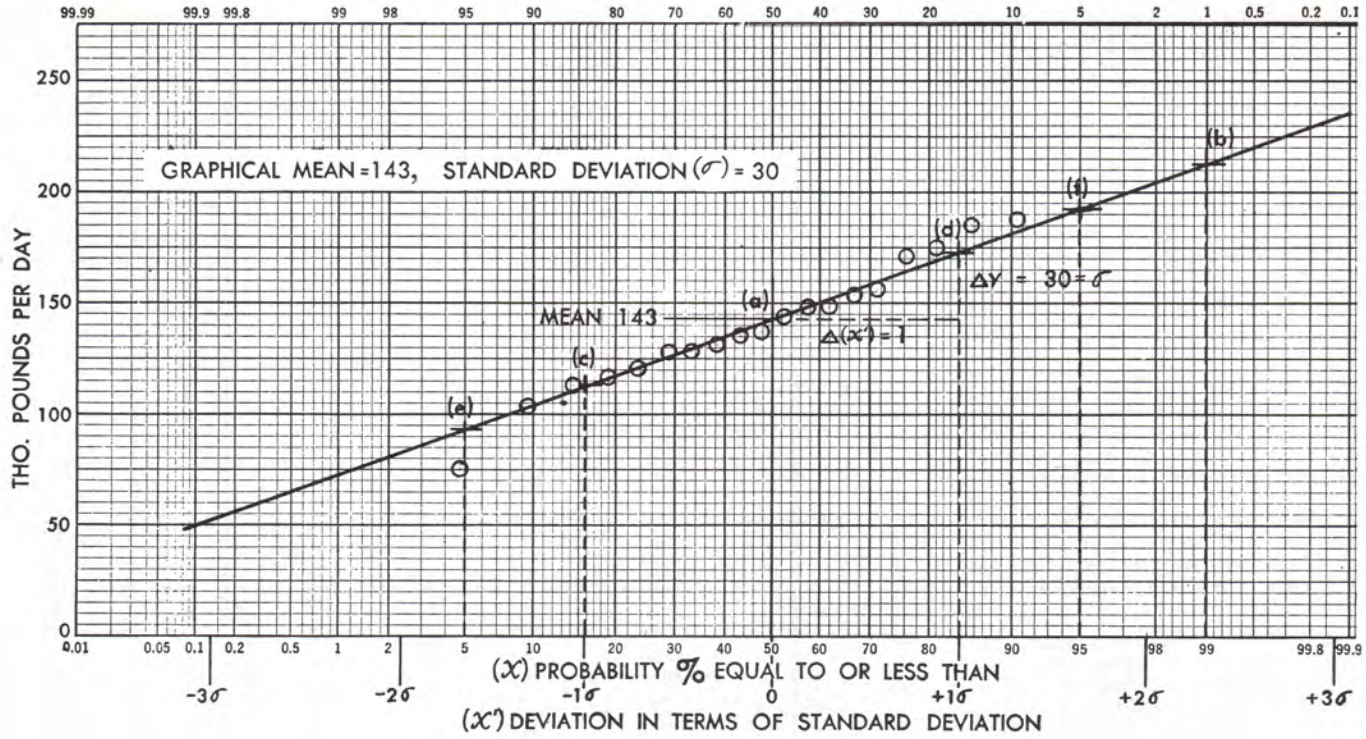


Fig. 6—Suspended Solids—Raw Sewage
(Buffalo Sewer Authority)

divisions of the group intervals, which corresponds to the serial number (m) as shown in Column 3. The plotting position (Column 4) on the X-axis of the probability paper is m/(n+1), expressed as per cent, where *n* is 500, the total number of values. The coordinates (Y) Column 1 and (X) Column 4 are then plotted on probability paper as shown in Fig. 7. The straight line formed summarizing the entire 500 values indicates a normal distribution, with a mean located at (a) of 63.6 ppm. and a slope or standard deviation from the mean of 11.4 ppm. The probability of any range, or of not exceeding any magnitude of measurement can be read directly from the graph. For example, we expect: not more than 1 per cent to exceed 90 ppm. located at (b); 1 per cent less than 37.3 ppm. located at (c); and 50 per cent within the range 56.0 to 71.3 ppm. about the mean between (d) and (e).

Reproducing a Series of Data

There are many occasions when we wish to reproduce graphically on probability paper a series of data from a published summary which does not tabulate each individual value. For example, if we are given the mean, the number of measurements and the standard deviation from the mean, it is a very simple operation to reconstruct on probability paper the expected normal distribution of the individual values of the series. This is just the reverse process of plotting a series of given data. For example, we are given the summary of a series of B.O.D. determinations, the number of measurements (n) 24, the mean ($\overline{Y}$) of 10.0 ppm., and the standard deviation (σ) from the mean of 1.57 ppm. The series is reproduced on Fig. 8 as follows:

Locate the mean on the Y-axis and draw a horizontal line to intersect with the vertical line, extended from 50 per cent on the X-axis at (a). This is the midpoint of our distribution. From this point lay off a slope $\Delta y/\Delta X'$ equal to the standard deviation 1.57, locating point (b). Draw a straight line from point (b) through the mean (a). This line represents the most probable distribution of the series of 24 individual measurements. If the original distribution was truly normal then the most probable values of each measurement would be located where m/(n+1) or m/(24+1) expressed in per cent extended vertically cut the line.

With this reconstruction we now can make any interpretations concerning confidence and range or probability of not exceeding any magnitude of measurement just as though we had been given the original data.

Thus far we have dealt with data that are normal, symmetrical about the central value and plot as straight lines. In the next article we shall deal with *skewed* distributions, such as selected extreme values, floods, droughts, and storm rainfall intensities.

Table 2
HARDNESS OF SOFTENED WATER SUPPLY—1949

Group Hardness ppm (Y) (1)	Frequency (2)	Summation of Frequency (m) (3)	Probability (%) equal to or less than [m/(n+1) as %] (X) (4)
30 —34	2	2	0.4
35 —39	6	8	1.60
40 —44	16	24	4.8
45 —49	32	56	11.2
50.0—54	53	109	21.8
55.0—59	74	183	36.5
60.0—64	87	270	53.8
65.0—69	85	355	70.9
70.0—74	66	421	84.0
75.0—79	42	463	92.4
80.0—84	22	485	96.8
85.0—89	10	495	98.8
90.0—94	4	499	99.60
95.0—99	1	500	99.80
	n = 500		

III—Use of Skewed Probability Paper

This section will deal with skewed probability paper as applied to series of extreme values such as floods, droughts, or storm rainfall intensities. It is well known that extremes of hydrologic phenomena such as floods and droughts do not follow a normal symmetrical distribution, but rather are skewed (the more severe values deviate beyond the mean to a much greater extent than the less severe values deviate below it). Gumbel* developed a standard skewed distribution based upon the theory of largest values which when reduced to straight line form on special probability paper as suggested by Powell†, simplifies and facilitates the graphical solution of many statistical problems dealing with series of extreme values.

A comparison of the normal probability curve and Gumbel's standard skewed distribution of largest values is shown on Fig.

*Gumbel, E. J. "The Return Period of Flood Flows," Annals of Mathematical Statistics, *12*, 163 (1941)

†Powell, R. W. "A Simple Method of Estimating Flood Frequencies," Civil Engineering, *13*:105 (1943)

y. On Fig. 9 (a) is the normal probability curve, which is perfectly symmetrical, with the mean and mode coinciding at the midpoint. Fig. 9 (b) is Gumbel's skewed distribution, which is asymmetrical, rising sharply to the peak and then falling off gradually in a long tail to the right, with the mode to the left of the mean. As with the normal curve, the probability of events is represented by the area under segments of the curve, total area being unity or 100 per cent. Gumbel has given these areas for increments of the curve and by summing from left to right a probability summation scale is obtained for the skewed distribution similar to that developed in the preceding article for normal probability paper.

Fig. 10 shows such a probability summation arranged to form a standard skewed probability paper applicable to series of extreme values. The ordinate (Y) is a linear scale assigned to the observed extreme values. Scale X is the probability summation scale giving the probability of a value *equal to or less than* a certain value Y. The X′ scale is introduced for linearity and as an aid in fitting a straight line, which will be illustrated later. T_x scale at the top is the "Return Period." If the observations are equally spaced in time and the unit of time is taken as the interval between two successive observations, the Return Period, T_x, is defined as the mean number of observations, or average time, necessary to obtain once a value equal to or greater than a certain value Y.

$$T_x = \left(\frac{1}{1-X}\right)$$

Maximum and Minimum Series

Gumbel's skewed probability paper is applicable to series of minimum values as well as maximum values providing the data are arranged in order of severity rather than magnitude, starting with the least severe value and ending with the most severe. In dealing with maximum series, such as floods, the least severe value would be the smallest observed flood, and the most severe value would be the largest observed flood. However, in dealing with minimum data, such as droughts expressed in cfs, the least severe drought would be the largest cfs among the observed droughts and the most severe drought would be the smallest cfs among the observed values, just the reverse of floods. This places the most severe flood and the most severe drought in the same position in terms of probability of severity. Thus by the simple technique of arranging maximum or minimum data in order of *severity* we place them in order of their position on the probability summation scale of the skewed probability paper.

Series of data that are truly extreme values will plot as a straight line on the skewed probability paper which facilitates interpolation and extrapolation. From such a straight line, probability of any severity can readily be determined graphically.

The Plotting Position

In the previous article the problem of the plotting position on the probability summation scale was discussed recommending $m/(n+1)$. This simple relation may also be employed for series of extreme values where m is now the serial number of data arranged in order of severity and n is the total number of values to be plotted.

Gumbel, however, has suggested a refinement which is recommended especially when dealing with a small series of say less than 20 observations. He develops the probability of the most probable least severe and most probable most severe value as the plotting position of the two extreme observations m_1 and m_n; and calculates the remaining (n-2) plotting positions by dividing the intervening interval into (n-1) equal units.

The plotting position for the least severe and the most severe values are complex functions of n, the number of observations,

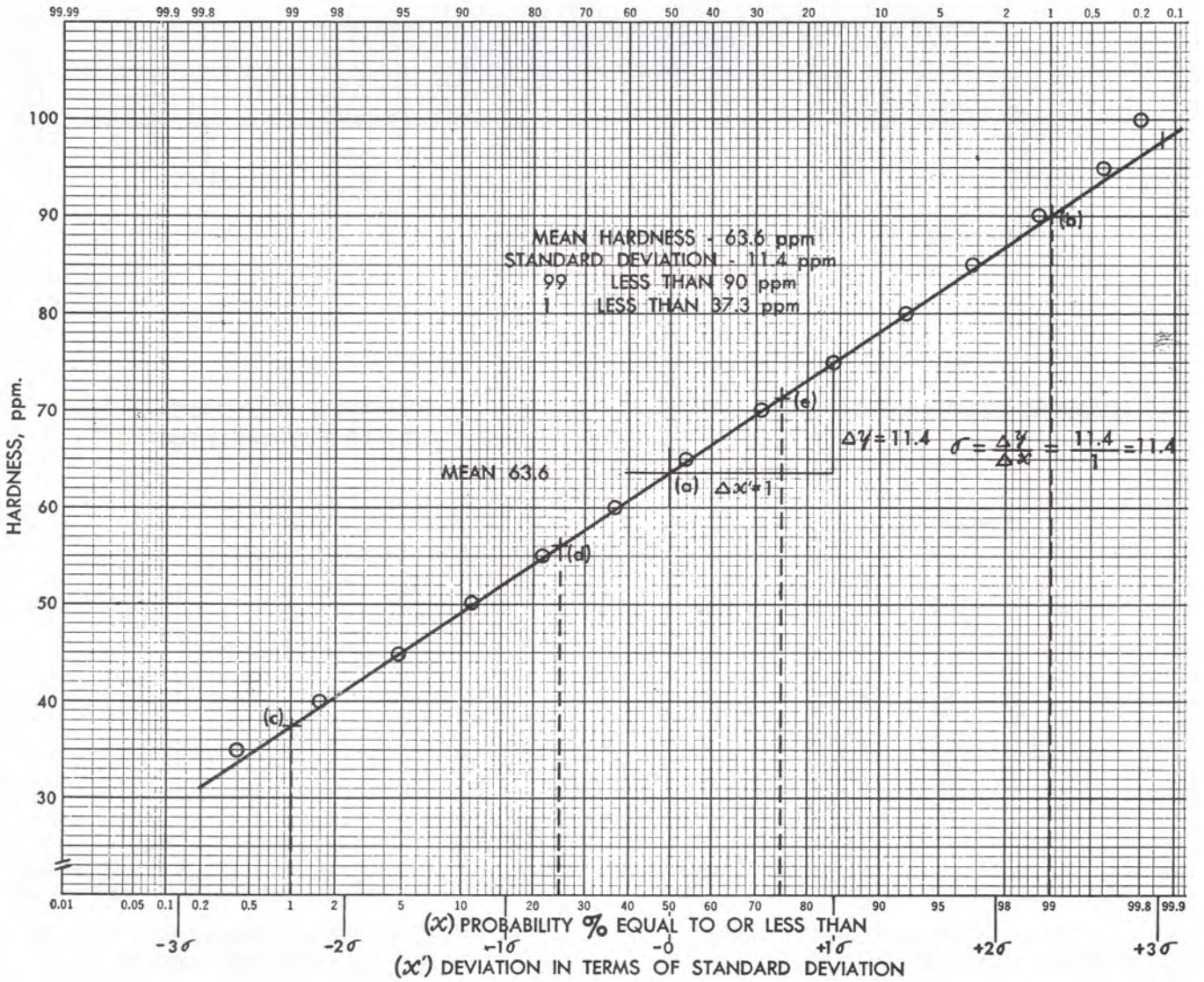


Fig. 7—Variation in Hardness in Softened Water Supply Based on 500 Determinations

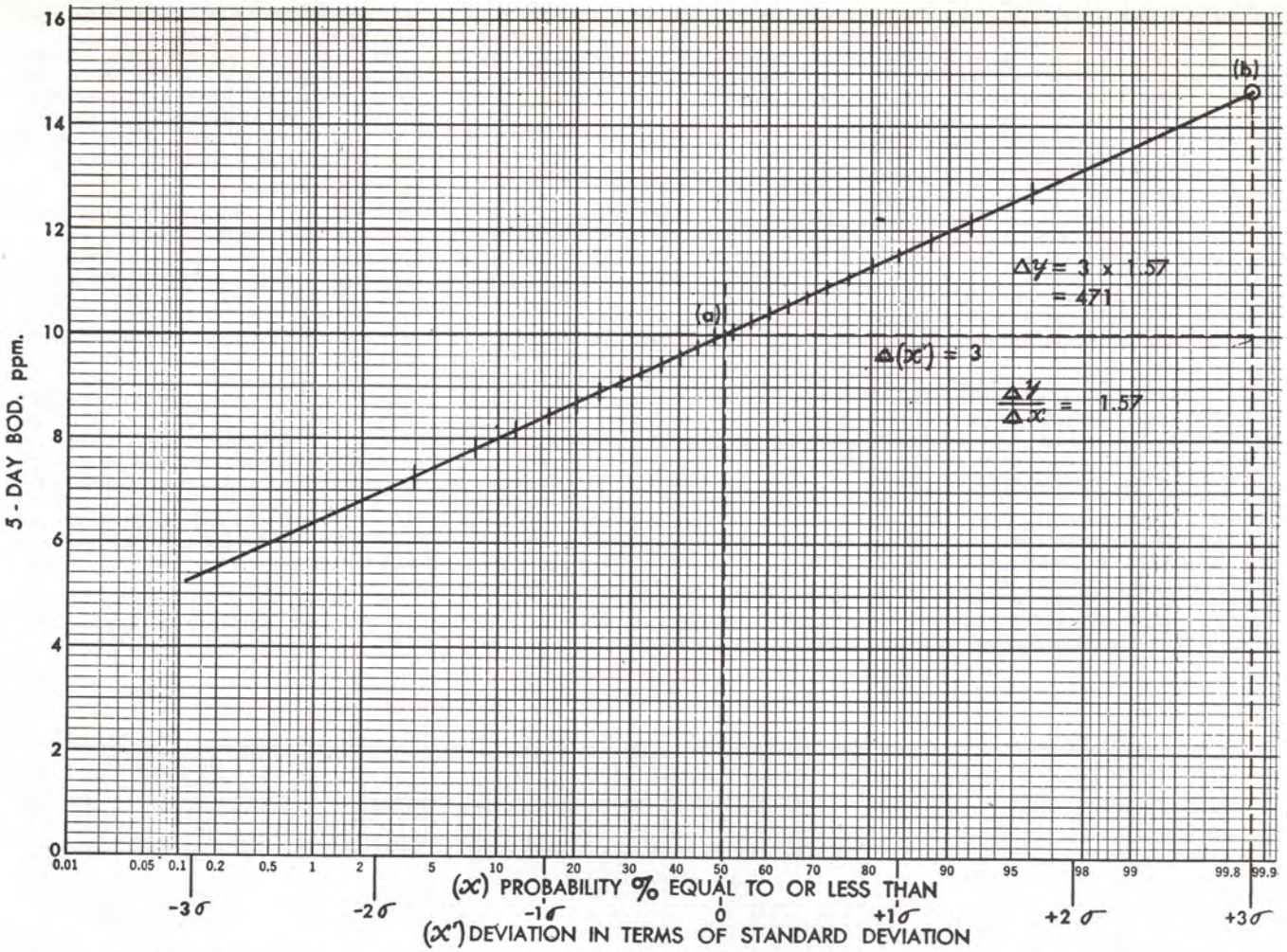


Fig. 8—Reproduction of Distribution of a Series of B.O.D. Determinations from Number, Mean, and Standard Deviations

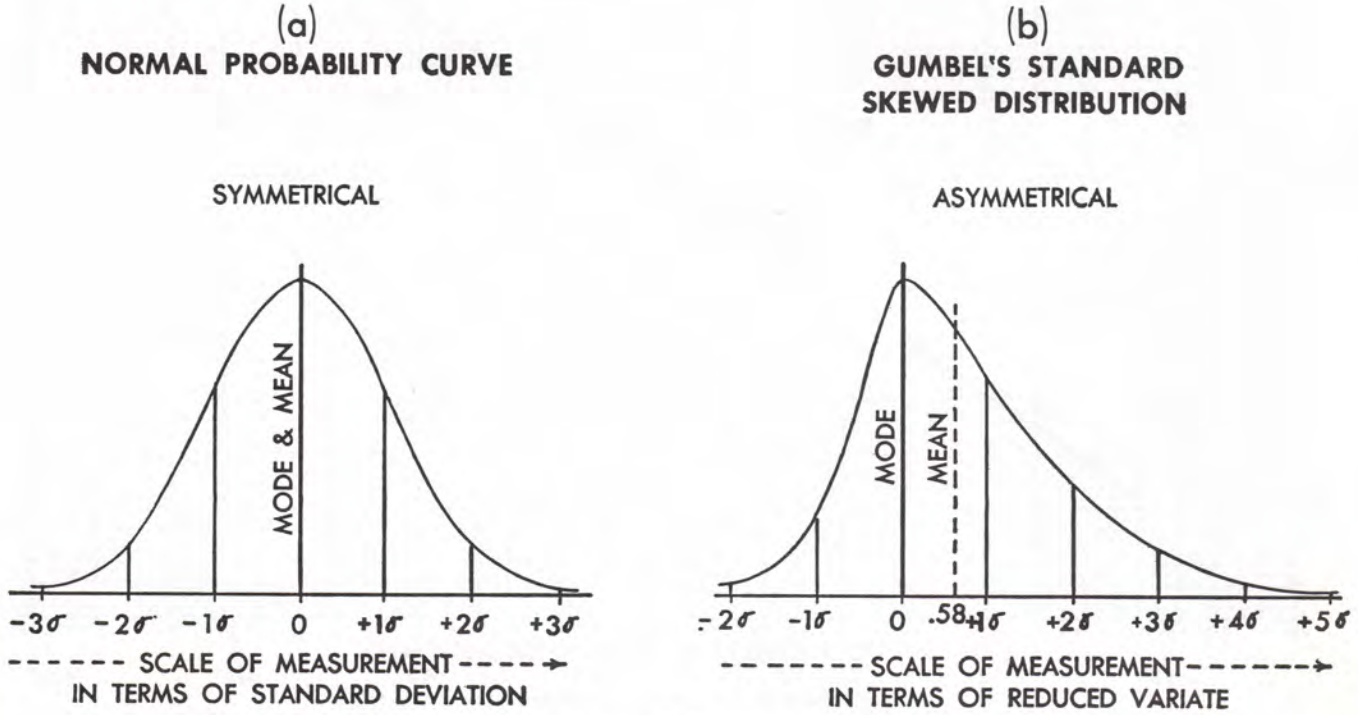


Fig. 9—Comparison of the Normal Probability Curve and Gumbel's Standard Skewed Distribution of Largest Values

Table 3

MAXIMUM ANNUAL FLOOD

Highest Peak Discharge In Each Water Year

Clark Fork Below Missoula, Mont.

Year	Discharge c.f.s.	Floods Arranged in order of Severity (Y)	Serial Number	Plotting Position (X)
(1)	(2)	(3)	(4)	(5)
1930	17,500	10,400	1	.0373
1931	12,200	11,700	2	.0879
1932	25,000	12,200	3	.1385
1933	36,800	14,200	4	.1892
1934	25,700	14,400	5	.2398
1935	20,200	17,500	6	.2904
1936	26,700	19,100	7	.3410
1937	11,700	19,700	8	.3917
1938	35,700	20,200	9	.4423
1939	22,000	22,000	10	.4929
1940	14,200	25,000	11	.5435
1941	10,400	25,700	12	.5941
1942	30,500	26,700	13	.6448
1943	33,200	30,500	14	.6954
1944	14,400	33,200	15	.7460
1945	19,100	35,700	16	.7966
1946	19,700	36,800	17	.8473
1947	45,900	45,900	18	.8979
1948	52,800	52,800	19	.9485

and for convenience may be read directly from Fig. 11. Only slight difference in the straight line plot on the skewed probability paper results between plotting positions determined from m/(n+1) or from Gumbel's refinement, except with a small number of observations.

Plotting Series of Extreme Values

In summary, the procedure for plotting series of extreme values, maximum or minimum, on skewed probability paper is as follows:

1. Arrange the data in order of severity commencing with the least severe.

2. Assign a serial number "*m*" to each of the *n* values, 1, 2, 3, . . . n.

3. Compute the plotting position of each serial value, giving the probability equal to less severe than, for each value either by m/(n+1) or by Gumbel's refinement.

4. Plot the points. The scale by which the observed measurements were made is laid off on the linear Y-axis; the computed plotting positions are the probability summation scale on the X-axis.

5. If a straight line is indicated by the plotted points it may be sketched graphically or it may be fitted by more precise means. Usually a graphical sketching is adequate.

The procedure and interpretations are best illustrated by specific applications which follow.

Flood Data

For purposes of illustrating a series of maximum values, the annual flood,* the highest peak discharge in each water year, is used. Table 3 represents such a series for Clark Fork below Missoula, Mont., drainage area 8,690 square miles, for the period 1930-1948. The plotting positions of Column 5 are in accordance with Gumbel's refinement obtained as follows:

Position of the most severe flood obtained from most severe curve, Fig. 11	.9485
Position for the least severe flood read from least severe curve, Fig. 11	.0373
Interval between least severe and most severe	.9112
Increment for successive points $\frac{.9112}{(19-1)}$	.050622

*Annual one-day flood, largest 24-hour average discharge among 365 days of each year, also develops a series of maximum values.

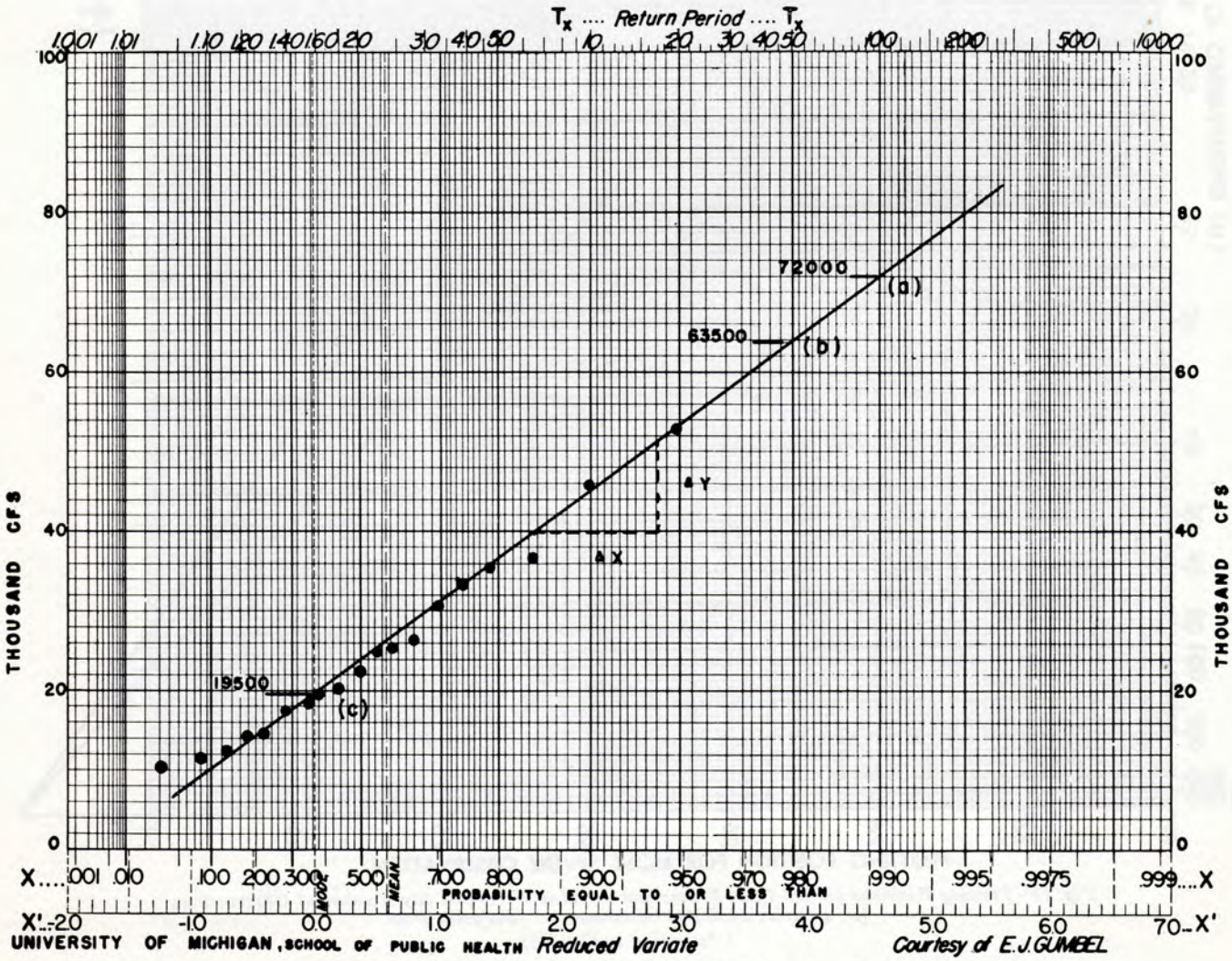


Fig. 10—Probability Annual Flood Discharge Clark Fork Below Missoula, Mont.

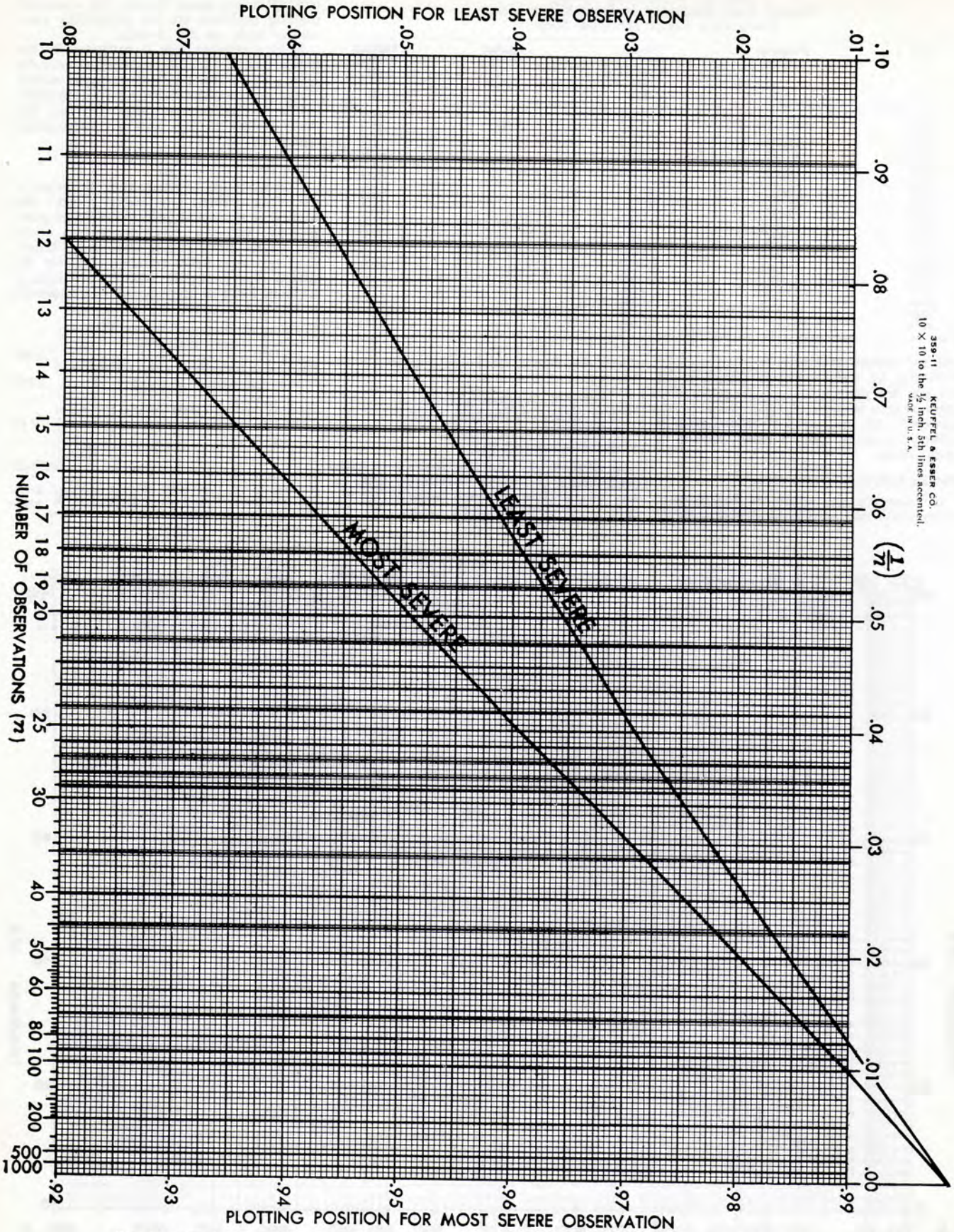


Fig. 11—Plotting Positions for the Least Severe and the Most Severe of any Number of Observations for a Series of Either Maximum or Minimum Data (According to Gumbel)

Thus the position of the least severe flood is at .0373, the second least severe flood at (.0373+.0506) or .0879, the third least severe at (.0879+.0506) or .1385, etc., adding .0506 for each successive value until the 19th or most severe value is reached with a plotting position .9485.

Column 3 and Column 5 of Table 3 provide the coordinates for plotting the data on skewed probability paper as shown on Fig. 10. A straight line is indicated as represented by the solid line fitted through the data. Thus, the 19-year record, while short, is in accord with the theory of extreme values and the fitted line may be accepted as a representative sample of flood expectancy for this river and probability of flood severity may be read directly from the line extrapolated beyond the record.

Since the mean return period interval between observations is a year, it is customary to speak of flood severity as equaled or exceeded in the long-run on the average once in 100 years as the 100-year flood. For the Clark Fork the 100-year flood is located by extending vertically from 100 on the T_x scale to intersect the distribuition line at (a), reading on the Y-scale 72,000 cfs. Similarly, in the long-run on the average once in 50 years a flood of 63,500 cfs. or greater may be expected, located at (b). The most probable flood value expected in any year is located where the mode cuts the distribution line, in this instance at (c) as 19,500 cfs.

Table 4

MINIMUM MONTHLY AVERAGE DROUGHT
Summer-Fall Period, May Through November, Ocmulgee River at Macon, Ga.

CFS. Per Sq. Mi. in order of Magnitude (Y) (1)	Serial Number (2)	Plotting Position (X) (3)
.954	1	.0373
.650	2	.0879
.485	3	.1385
.447	4	.1892
.446	5	.2398
.412	6	.2904
.412	7	.3410
.380	8	.3917
.377	9	.4423
.374	10	.4929
.360	11	.5435
.340	12	.5941
.335	13	.6448
.314	14	.6954
.298	15	.7460
.264	16	.7966
.228	17	.8473
.218	18	.8979
.114	19	.9485

The slope of the line on skewed probability paper is a measure of the degree of variation among the annual flood values. This slope may be obtained graphically by scaling any $\frac{\Delta Y}{\Delta X'}$ developed by the distribution line, employing the reduced variate scale for $\Delta X'$. If the flood flows are expressed in yields, cfs. per square mile of tributary drainage area, slopes for flood distributions for different drainage areas are directly comparable. In this example $\frac{\Delta Y}{\Delta X'}$ 11,400 in terms of cfs., or 1.31 in terms of yield. In a similar manner, if the Most Probable Flood (located by the Mode, $X'=O$) is also expressed as a yield, cfs. per square mile of tributary drainage area (in this example 19,500/8,690 or 2.245), a comparative measure of flood potential is obtained. The slope and the Most Probable Flood (MPF) expressed in terms of yield are two significant indices for comparative study of flood characteristics. These indices also afford an excellent summary, as the complete distribution line can be reconstructed simply by locating the MPF on the modal line and from this point lay off the slope and draw the distribution line.

Drought Data

For purposes of developing a series of minimum values, minimum monthly* average drought flows occurring during the summer-fall period, May through November, of each year have been selected from

*Minimum daily average or minimum weekly average discharge data also develop series of minimum extreme values.

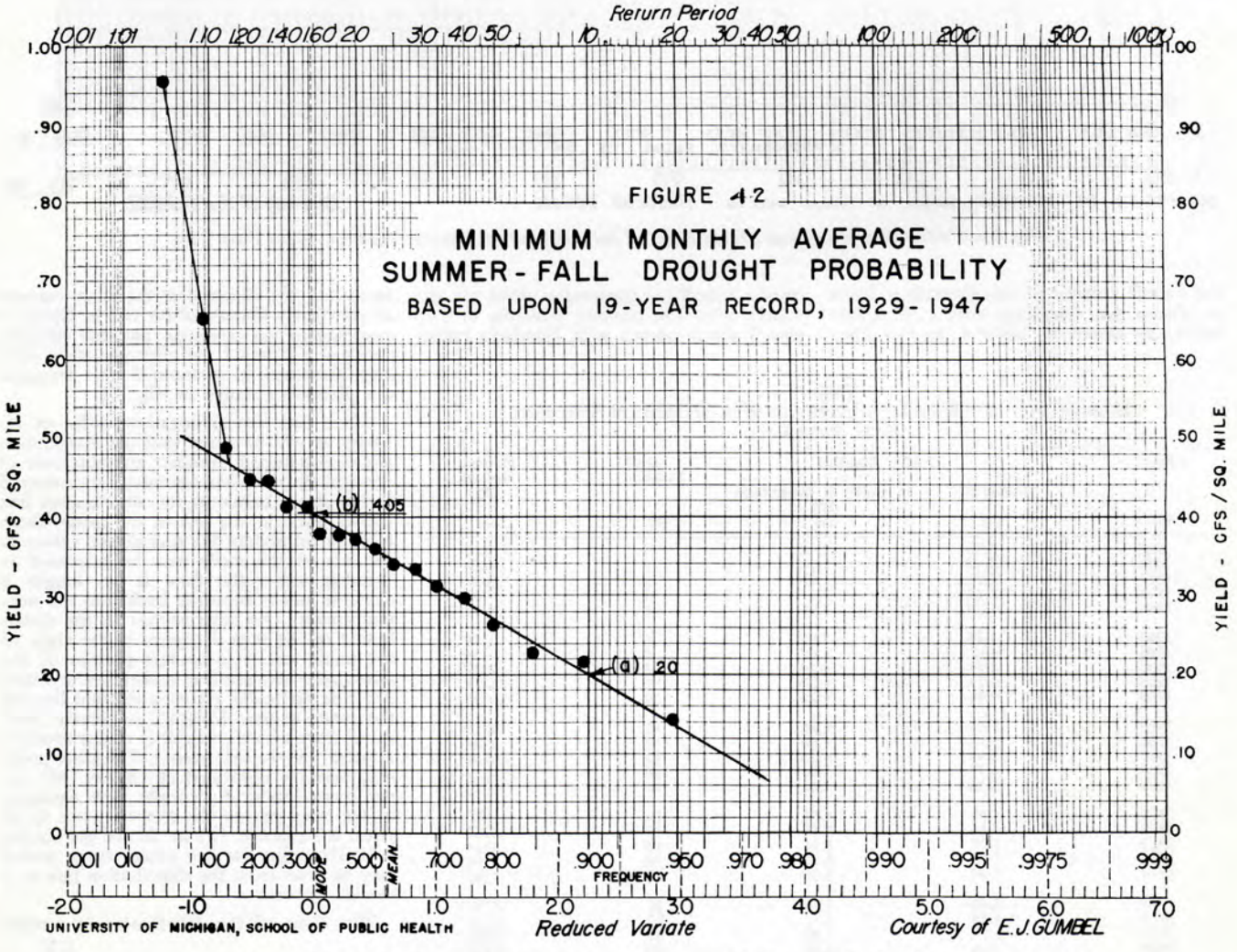


Fig. 12—Minimum Monthly Average Summer-Fall Drought Probability
(Based upon 19-year record, 1929-47.)

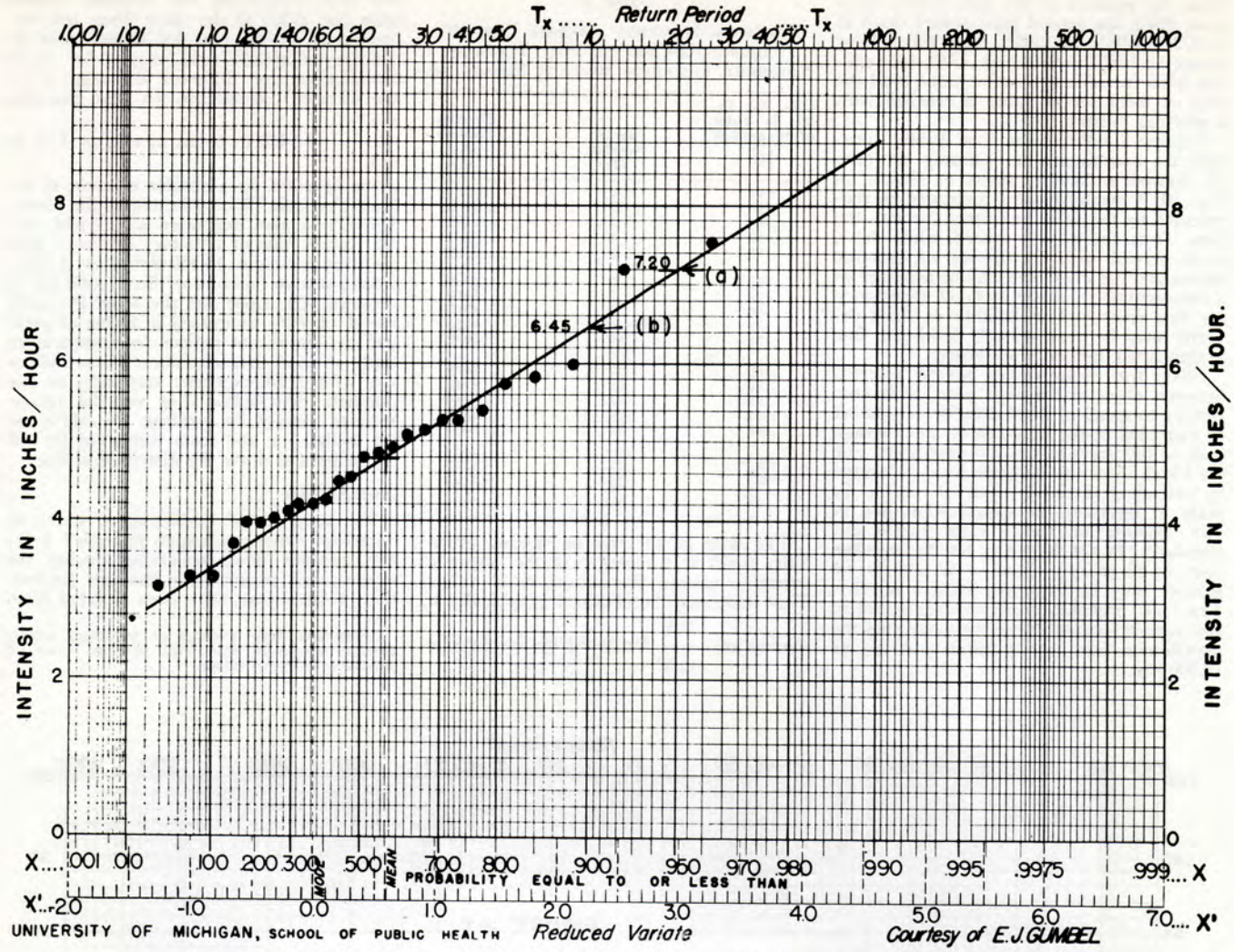


Fig. 13—Probability Maximum Storm Rainfall Intensity for 10 Minute Duration, New York City

the runoff records of the Ocmulgee River at Macon, Ga. (drainage area 2,240 square miles) as shown in Table 4. In this illustration, runoff is expressed as yield, cfs. per square mile. The plotting positions in Column 3 are in accord with Gumbel's refinement and are obtained in the same manner as previously illustrated for floods. Using as coordinates the minimum monthly drought flows of Column 1 and corresponding plotting positions in Column 3 the droughts are plotted as shown in Fig. 12.

Table 5
MAXIMUM STORM RAINFALL INTENSITY FOR 10 MINUTE DURATION
NEW YORK CITY

Year	10 Min. Duration		Serial Number	Plotting Position (X)
	Inches/Hr.	in order of Magnitude (Y)		
(1)	(2)	(3)	(4)	(5)
1895	3.96	3.18	1	0.0282
1896	5.10	3.30	2	.0656
1897	3.72	3.30	3	.1029
1898	3.18	3.72	4	.1403
1899	4.20	3.96	5	.1777
1900	3.30	3.96	6	.2150
1901	4.92	4.02	7	.2524
1902	5.16	4.14	8	.2897
1903	4.14	4.20	9	.3271
1904	4.86	4.20	10	.3645
1905	7.56	4.26	11	.4018
1906	5.82	4.50	12	.4392
1907	3.96	4.56	13	.4766
1908	3.30	4.80	14	.5139
1909	5.28	4.86	15	.5513
1910	6.00	4.92	16	.5887
1911	4.56	5.10	17	.6260
1912	5.28	5.16	18	.6634
1913	7.20	5.28	19	.7008
1914	4.80	5.28	20	.7381
1915	5.40	5.40	21	.7755
1916	4.50	5.76	22	.8128
1917	5.76	5.82	23	.8502
1918	4.02	6.00	24	.8876
1919	4.20	7.20	25	.9249
1920	4.26	7.56	26	.9623

Since data are arranged in order of severity, it will be noted that the distribution slopes downward. Another characteristic of drought distributions commonly encountered is the sharp break in the distribution line taking place at the left of the mode. Such a break formed by the less severe values to the left of the mode may be regarded as defining when the flow is no longer a drought and is blending back into the normal runoff. Our interest lies in the distribution of the true droughts to the right of the mode and it is to this portion of the data that a straight line is sketched or fitted. For this particular record a straight line for the more severe droughts is clearly indicated, from which probability of any drought severity may be determined. The most probable drought is located at (b) as .405 cfs. per square mile. A drought flow expected in the long run on the average once in 10 years is located at (a) as .20 cfs. per square mile. Droughts for any other return period may be read from the distribution line in a similar manner.

The slope of the distribution is readily obtained graphically by scaling $\frac{\Delta Y}{\Delta X'}$ or 0.092.

Storm Rainfall Intensity

A good example of a series of maximum values is the maximum rainfall intensity as an average for a given duration period occurring at the center of the worst storm each year in a localized geographical region. Table 5 represents such a series of data for 26 years from 11 rain gaging station in New York City, selecting the maximum intensity for a 10 minute duration. The plotting positions (Column 5) for the 26 values, again, are in accord with Gumbel's refinement as illustrated for floods. The series plots as a straight line throughout, as shown on Figure 13. From the distribution line the probability of any intensity can be determined. Thus on the average in the long run an intensity, as an average for 10 minute duration, of 7.20 inches per hour or greater may be expected once in 20 years, located at (a); and 6.45 inches per hour once in 10 years. Intensities for other duration periods form similar straight line distributions, and from them a rational set of storm rainfall intensity curves can readily be constructed to serve as a basis for storm water drainage design.

Water Supply Storage

Fig. 14 illustrates an application of skewed probability paper to a storage problem. The maximum annual storage required each year to meet a demand of 500 mgd. was taken from a mass curve of net run-off from the Schoharie-Esopus basin for the period 1903-1945. Since drought severity is reflected in magnitude of storage required, the storage values are arranged in ascending order of magnitude, as shown in Table 6. Plotting positions for the 43 values are in accord with Gumbel's refinement and are obtained in the same manner as illustrated for floods. The storage values associated with true droughts are those points located to the right of the mode and the distribution line should be sketched or fitted through these values. The small storages associated with the less severe droughts to the left of the mode usually fall below the distribution line. The probability of any storage requirement to insure a water supply demand of 500 mgd. can be read directly from the distribution line. Thus, to protect against drought severity expected on the average in the long run once in 50 years, it would be necessary to provide storage of 100 billion gallons, located at (a) on Figure 14. This would provide storage sufficient to meet demand on 98 per cent of the years. Similarly, for drought severity expected once in 20 years the storage required would be 85 billion gallons, located at (b), providing for 95 per cent of the years.

In some storage situations the demand is set so high that reservoirs do not refill on the average each year. Under these conditions the unit of time may become 2 or 3 years and one maximum storage is selected from the mass curve for each such interval. Data are handled in the same manner except that the return period is in the new unit of time and is no longer directly in years.

IV—Evaluation of Bacterial Density

This section will deal with logarithmic probability paper as applied to the evaluation of bacterial density. Certain types of data such as bacterial density determined by the decimal dilution method, are logarithmically normal. Such a distribution will plot as a straight line on normal probability paper previously described providing the logarithms of the measurements are substituted for the original data. Log-probability paper avoids converting the data to logarithms by employing a logarithmic grid as the ordinate scale. The original data may then be plotted directly on the log-grid.

Table 6

WATER SUPPLY STORAGE REQUIREMENTS FOR A DEMAND OF 500 MGD.
Based Upon Mass Curve of Net Yield
Schoharie-Esopus Drainage Basin, 671 Sq. Mi.

Year (1)	Storage Required Billion Gallons (2)	in order of Magnitude (Y) (3)	Serial Number (4)	Plotting Position (X) (5)
1903	5	4	1	.0177
1904	32	5	2	.0405
1905	52	8	3	.0635
1906	40	9	4	.0861
1907	27	9	5	.1089
1908	60	14	6	.1317
1909	67	20	7	.1545
1910	58	21	8	.1773
1911	32	21	9	.2001
1912	40	27	10	.2229
1913	47	28	11	.2457
1914	76	28	12	.2685
1915	14	29	13	.2913
1916	21	32	14	.3141
1917	29	32	15	.3369
1918	49	33	16	.3597
1919	9	36	17	.3825
1920	21	39	18	.4053
1921	55	40	19	.4281
1922	60	40	20	.4509
1923	40	40	21	.4737
1924	33	45	22	.4965
1925	50	46	23	.5193
1926	50	47	24	.5421
1927	20	48	25	.5649
1928	45	49	26	.5877
1929	48	49	27	.6105
1930	94	50	28	.6333
1931	60	50	29	.6561
1932	61	52	30	.6789
1933	28	55	31	.7017
1934	39	55	32	.7245
1935	36	58	33	.7475
1936	70	60	34	.7701
1937	9	60	35	.7929
1938	8	60	36	.8157
1939	94	61	37	.8385
1940	46	67	38	.8613
1941	71	70	39	.8841
1942	28	71	40	.908
1943	49	76	41	.9305
1944	55	94	42	.954
1945	4	94	43	.977

Fig. 15 illustrates a logarithmically normal distribution plotted on log-probability paper. The X and X′ scales are identical with those previously described for normal probability paper; X-scale at the bottom is probability (in per cent) equal to or less than or probability of not exceeding, X′-scale at the top is deviation from the mean in terms of standard deviation. The left ordinate scale, Y, is a logarithmic grid assigned to the scale of measurement of the original data; Y′-scale at the right is a linear grid on which logarithms of the original data are located.

Table 7

16 SAMPLES TAKEN SIMULTANEOUSLY FROM THE SAME SOURCE 10 PORTIONS IN EACH OF 3 DECIMAL DILUTIONS

MPN per ml in order of Magnitude Millions	Serial Number m	Plotting Position m/(n+1) as %
2.88	1	5.9
3.29	2	11.8
3.29	3	17.6
3.34	4	23.5
3.99	5	29.4
3.99	6	35.3
4.74	7	41.2
4.74	8	47.0
4.93	9	52.8
4.93	10	58.3
5.89	11	64.6
6.22	12	70.5
6.22	13	76.5
7.00	14	82.3
7.42	15	88.2
9.33	16	94.1

Plotting on Log-Probability Paper

The procedure in plotting on log-probability paper is similar to that employed for normal probability paper:

(1) Arrange the data in order of ascending magnitude.

(2) Assign a serial number "m" to each of the n values, 1, 2, 3, ... n.

(3) Compute the plotting position of each serial value, giving the probability equal to or less than for each value by the expression $m/(n+1)$ expressing the ratio as a percentage.

(4) The units of measurement of the original data are laid off on the logarithmic grid* as the Y-axis and the plotting positions are the probability scale on the X-axis. The points are plotted using these (X, Y) coordinates. (Table 7 shows the coordinates for the 16 values plotted on Fig. 15.)

In interpreting a plot on log-probability paper it is cautioned that while the original data are plotted on the Y scale, actually it is the corresponding Y′ values, the logarithms of the original data, that form the normal distribution. Hence the mean and the standard deviation are expressed in terms of logarithms. The subscript (Log)

*If more than two log cycles are needed, sheets of log-probability paper can be pasted together.

is employed to indicate that a logarithmically normal distribution is dealt with. The procedure for determining graphically the $\text{mean}_{(\text{Log})}$ and $\text{sigma}_{(\text{Log})}$ is identical with that described for normal probability paper except that Y′ scale is employed. Referring to Fig. 15

$$\text{Mean}_{(\text{Log})} = 6.687$$

located on the Y′ scale opposite (a) where 50 per cent cuts the distribution line. This is equivalent to the arithmetic mean of the logarithms of the original values.

The slope of the distribution line is the measure of variation of the data and provides the standard deviation in terms of logarithms.

$$\sigma_{(\text{Log})} = \frac{\Delta Y'}{\Delta X'} = .174$$

$\Delta Y'$ is measured on Y′ scale and $\Delta X'$ is measured on the X′ scale.

It is cautioned that in computing analytically ranges for any confidence it is essential first to deal in terms of logarithms. After the logarithms of the limits of any range are determined their anti-logs can be obtained and results expressed in terms of the original units of measurements. The advantage of log-probability paper is that these conversions are automatically accomplished. Again referring to Fig. 15, the $\text{mean}_{(\text{Log})}$ 6.687, is converted to units of the original measurement simply by reading directly on the Y scale, giving a value 4,860,000 per ml. Similarly the range for $\text{mean}_{(\text{Log})} \pm 2\sigma_{(\text{Log})}$ in terms of logarithms is

$$6.687 \pm 2 \times .174 \qquad (6.339 \text{ to } 7.035)$$

located by (b) and (c) where -2σ and $+2\sigma$ intersect the distribution line, reading on the Y′ scale. The limits of this range converted to units of the original measurements are directly obtained by reading on the Y scale

(2,180,000 per ml to 10,820,000 per ml)*

Concepts of Mean Density and Distribution of Bacteria

It is first essential to introduce some basic concepts about bacterial density in a liquid medium. If it were possible to count the entire bacterial population in the entire water supply, the true *Mean Density* (M.D.) would be the total count divided by the total volume. If less than the total volume were employed in counting, it would be found that the densities in the samples would vary about the true Mean Density, the smaller samples showing greater variation than the larger samples. Perfect admixture would require exactly the same number of bacteria in each and every unit volume; complete segregation, on the other hand, would require all cells to be located in one single unit volume. Such extreme distributions of bacteria among unit volumes of water do not exist. The actual distribution is random and would tend toward an equilibrium condition between these extremes, governed by the laws of chance.

Inherently, therefore, we expect variation in results from small samples even if they are all drawn from the same supply at the same instant. Fortunately, the nature of these expected variations can be defined and series of test results can be reduced to reliable estimates of the true Mean Density as the basic measure of bacterial quality.

Nature of Variation of MPN's

Three fundamental characteristics of the nature of variation in MPN's determined from test results employing multiple portions in three decimal dilutions are:

(1) The distribution is logarithmically normal, that is, series of MPN's plot as a straight line on log-probability paper.

(2) The estimate of the true *Mean Density* is located at the midpoint of the distribution at 50 per cent.

(3) The slope of the distribution line depends upon the number of portions employed; small number of portions steep

*Note, that the deviations in terms of the original units of measure (Y) are not equal about the mean; $2\sigma_{(\text{Log})}$ plus $\text{mean}_{(\text{Log})}$ produces a Y of 10,820,000 or 5,960,000 above the mean of 4,860,000 while $2\sigma_{(\text{Log})}$ minus $\text{mean}_{(\text{Log})}$ produces a Y of 2,680,000 only 2,680,000 below the mean. This is inherent in the skew of the *logarithmically* normal distribution.

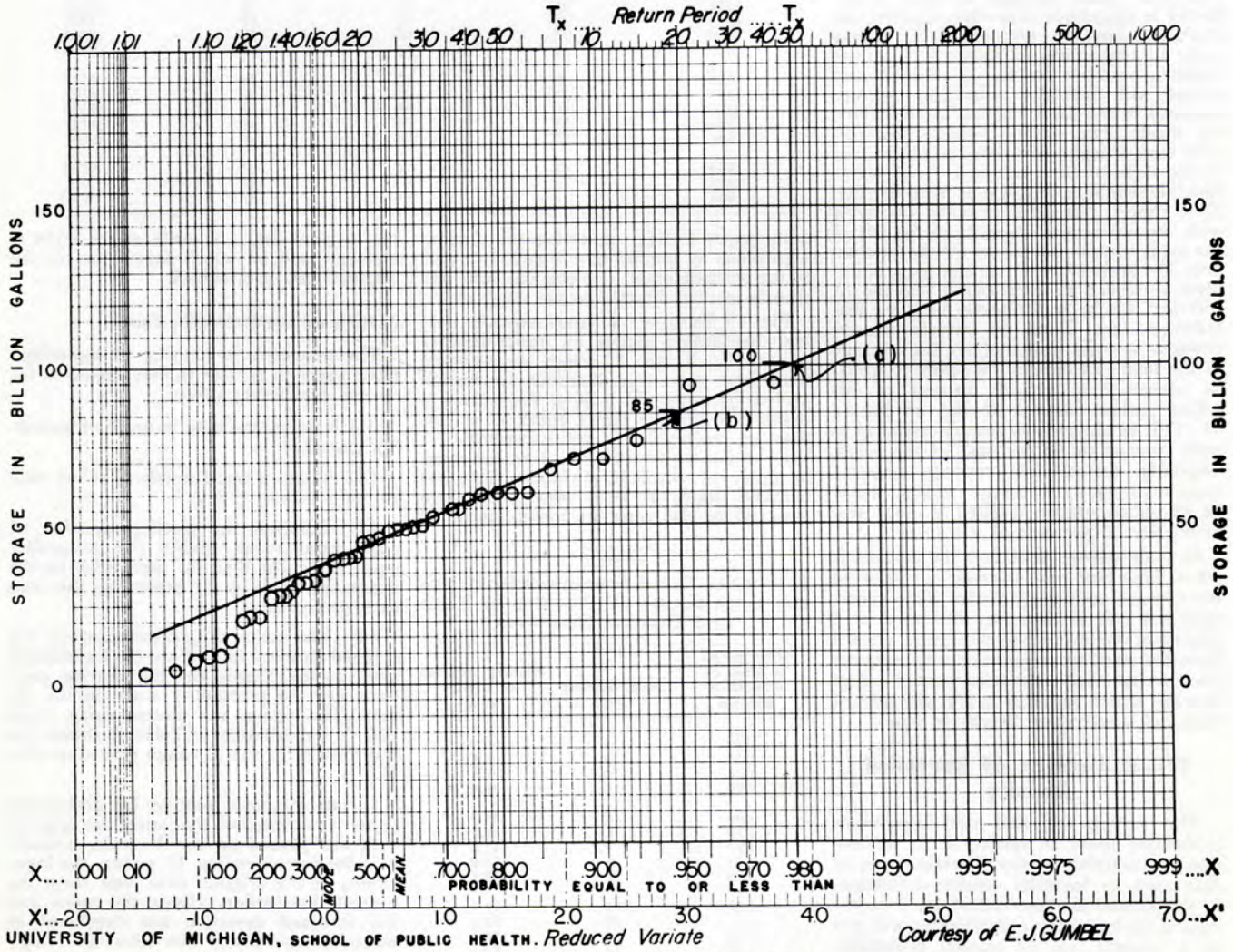


Fig. 14—Probability Maximum Storage Requirements for Water Supply Demand of 500 MGD. (Schoharie-Esopus Basin, 671 sq. mi.)

slope, wide variation; large number of portions flat slope, narrow variation.

These characteristics can be developed mathematically [1] and verified experimentally.

Figure 16 shows the results of theoretical probability distribution of possible results from *20 portions* in each of 3 decimal dilutions with a known true Mean Density of 1.5 bacteria per ml. Note:

(1) Almost a perfect straight line develops.

(2) The distribution line cuts 50 per cent at a density of 1.5 which is equal to the theoretical true Mean Density employed.

(3) The slope is relatively flat, with $\sigma_{(Log)} = 0.1227$.

Figure 17, likewise, shows the results of theoretical probability distribution of possible results with the same true Mean Density 1.5, but employing only *5 portions* in each of 3 decimal dilutions. Note again:

(1) A straight line develops.

(2) The distribution line cuts 50 per cent at the theoretical true Mean Density of 1.5.

(3) The slope is steeper with

$$\sigma_{(Log)} = 0.2454$$

More generally $\sigma_{(Log)}$, the slope, is related to the number of portions (n) as

$$\sigma_{(Log)} = \frac{.5487}{\sqrt{n}}$$

Graphically, the slope ($\sigma_{(Log)}$) for various values of (n), number of portions, is shown in Fig. 18. These are the standard distribution slopes expected purely by virtue of the test procedure, depending solely on the number of *portions* employed. All the distribution lines arise from water of the same Mean Density but the variation of individual MPN values can be expected to be greater and greater as the number of portions is decreased.

Hence to know the reliability of an MPN it is essential to know the number of *portions* employed in the test procedure.

These relationships can be verified experimentally by taking a set of samples simultaneously from the same source, thus insuring a constant Mean Density for all samples. Fig. 15, previously referred to, represents the distribution of MPN values for 16 samples taken from a controlled source in such a manner, each sample composed of *10 portions* in each of 3 decimal dilutions.

The slope of the line described by the plotted points is in remarkable agreement with the theoretical or standard slope expected when 10 portions are employed, namely $\sigma_{(Log)} = 0.1735$. As the number of samples increases the experimental slope approaches the theoretical standard slope as a limit. This holds of course only if there is no change in the true Mean Density during the period when samples are drawn.

Figure 19 is the distribution of MPN's of a set of 16 samples taken simultaneously from a raw water supply, employing *3 portions* in each of 3 decimal dilutions. When only 3 portions in each of 3 decimal dilutions are employed the number of possible outcomes is less and the distribution is less continuous in character, but the points describe a line in close agreement with the theoretical or standard slope expected for *3 portions*, namely, $\sigma_{(Log)} = 0.32$.

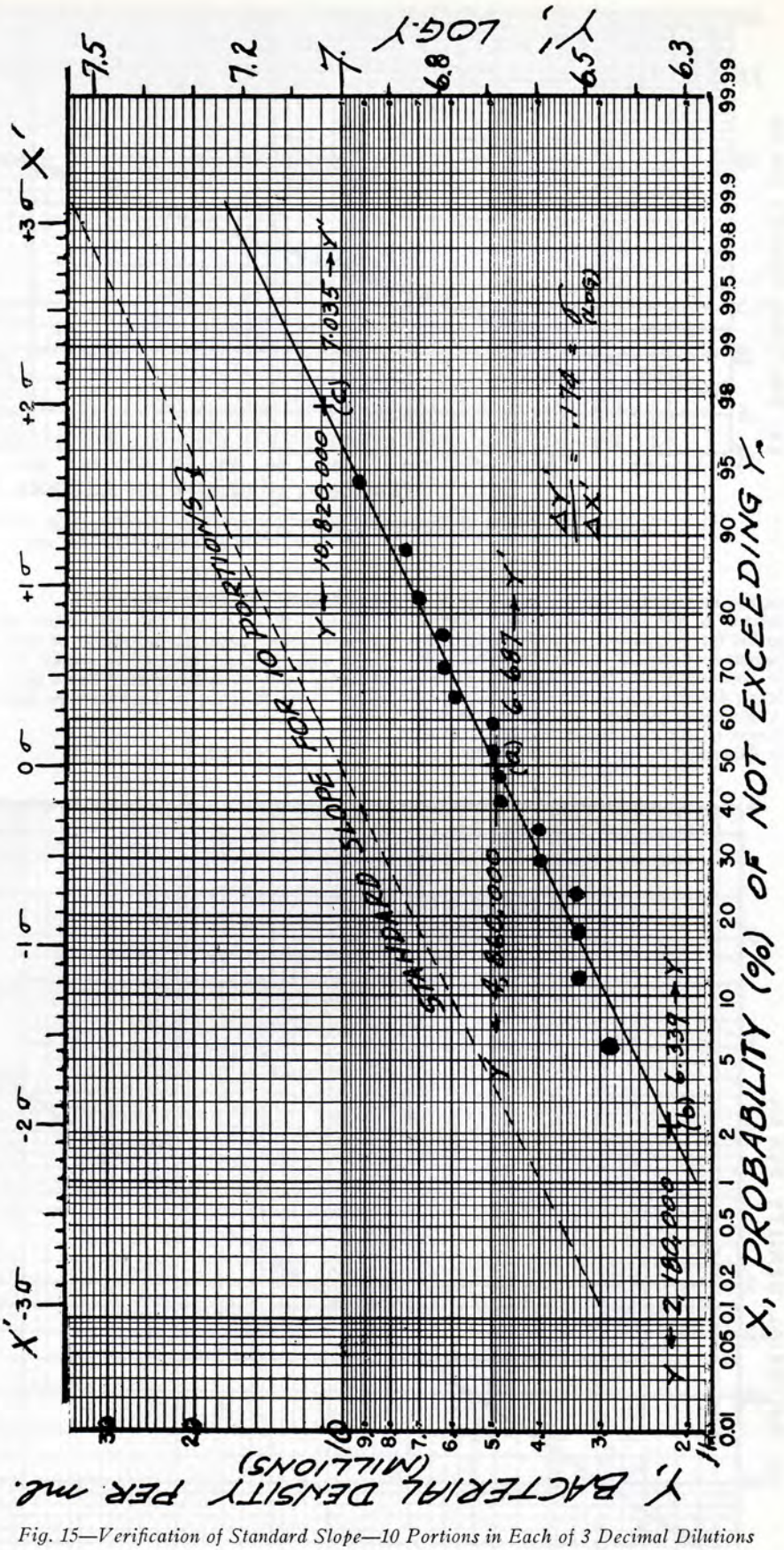


Fig. 15—Verification of Standard Slope—10 Portions in Each of 3 Decimal Dilutions

If samples are taken over a time interval, such as daily routine samples over a period of a month, true Mean Density undoubtedly would not remain constant and hence the daily MPN's would be drawn from water of changing quality. Under these conditions of changing quality, the distribution line described by the individual MPN's will be steeper than that where

(1) *Proceedings Inservice Training Course in Sewage and Industrial Waste Disposal*, 1949, School of Public Health, University of Michigan.

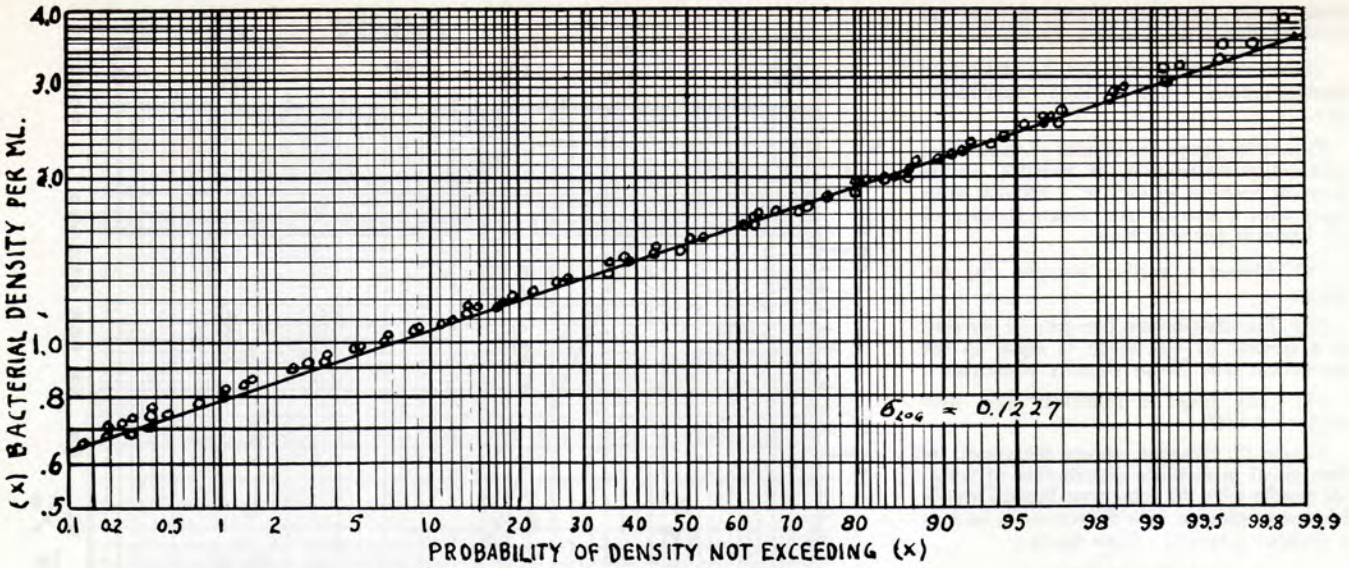


Fig. 16—Distribution of Bacterial Density Expected for Sample of 20 Portions Each of 3 Decimal Dilutions at True Mean Density of 1.5 per ml.

quality remained constant. Hence the comparison of the actual slope with that expected for the number of portions employed indicates at once if quality remained stable or was subject to real changes in level of Mean Density during the sampling period. A graphical method for determining the extent of such a change in Mean Density is illustrated in the following examples.

Evaluation of Raw Water Supply Data

Figure 20 represents the application of log-probability paper to evaluation of coliform density of the raw water supply for the City of Wyandotte, Mich.,* based upon daily routine samples of 3 portions in each of 3 decimal dilutions for the 71 day period July 22 to Sept. 30, 1950. A reasonably straight line distribution is described by the distribution of the 71 MPN values shown as heavy dots.

THE MEAN DENSITY—The graphical line "A" through the data cuts 50 per cent at $Mean_{(Log)}$ of 4.447 (read on Y'-scale or a

*Data courtesy George Hazey.

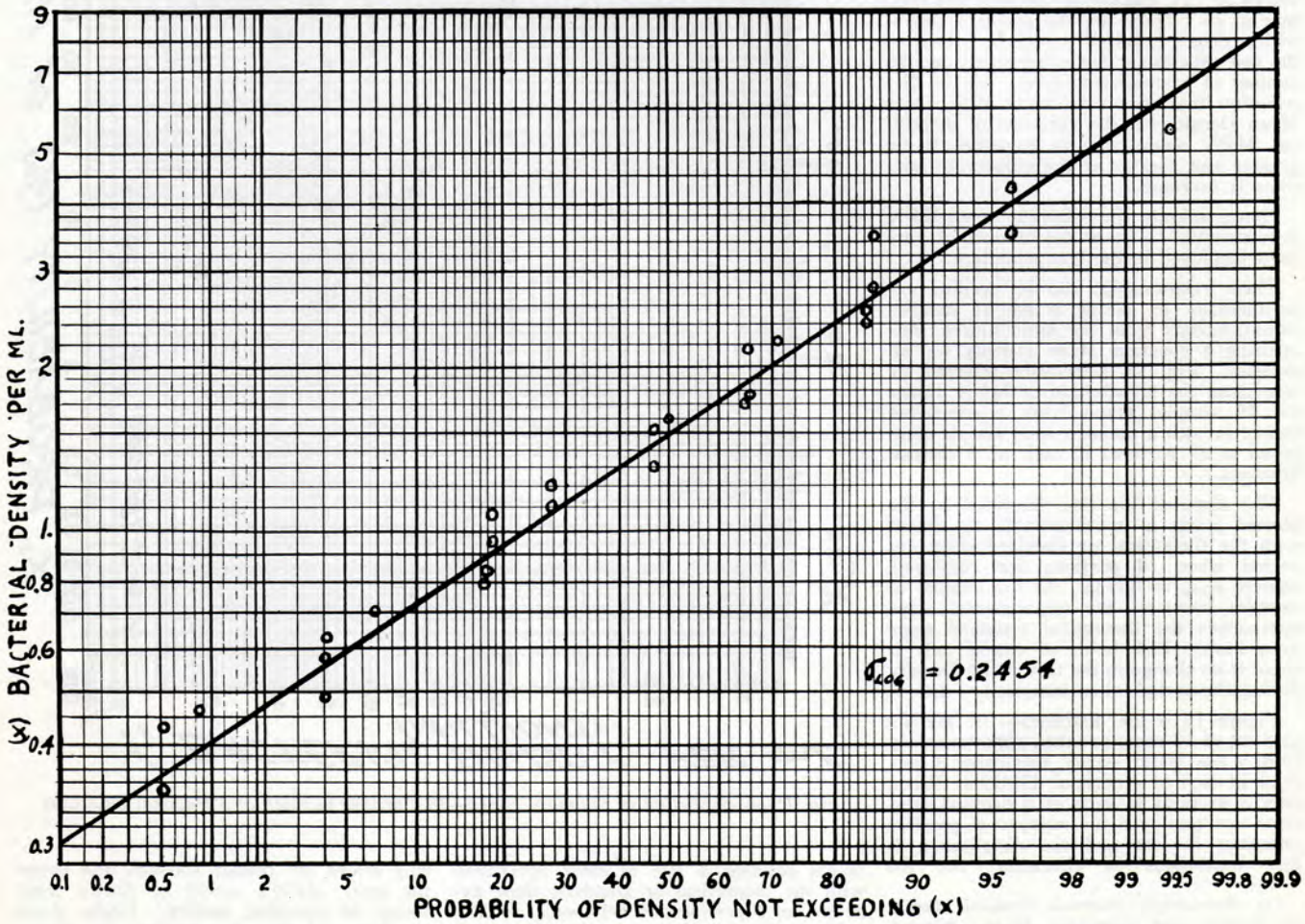


Fig. 17—Distribution of Bacterial Density Expected for Sample of 5 Portions Each of 3 Decimal Dilutions at True Mean Density of 1.5 per ml.

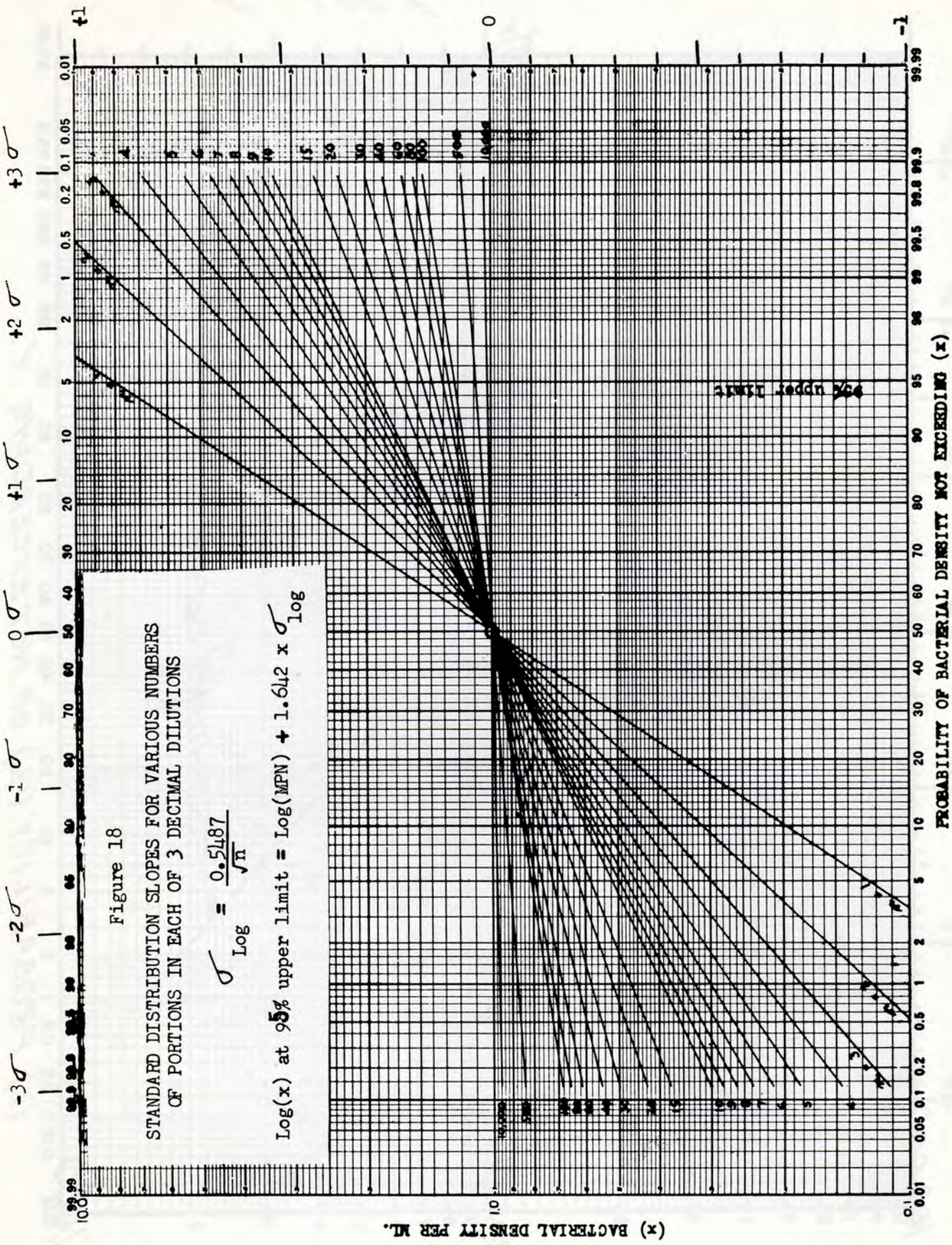


Fig. 18—Standard Distribution Slopes for Various Numbers of Portions in Each of 3 Decimal Dilutions

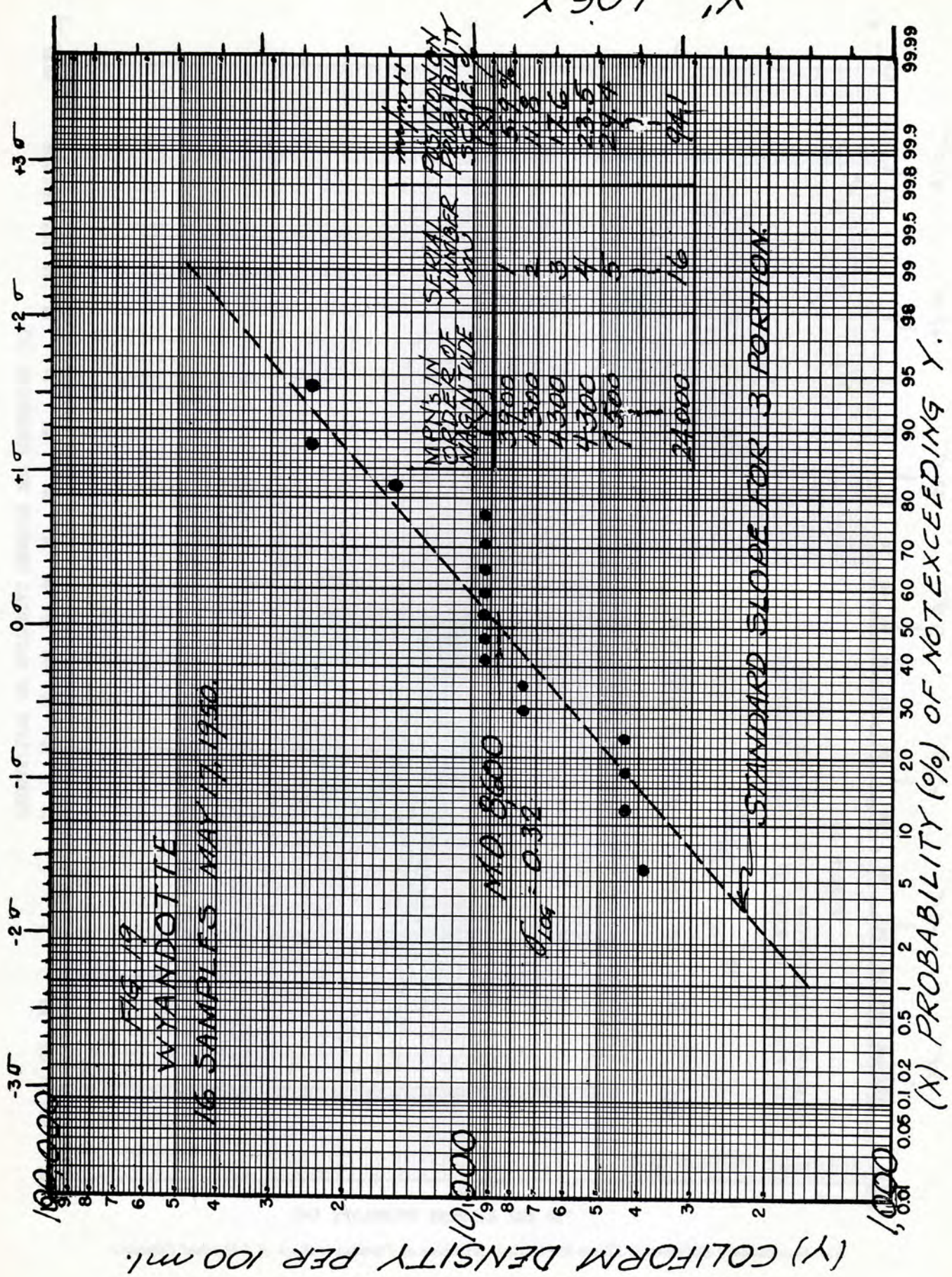


Fig. 19—Wyandotte, Mich.—16 Samples May 17, 1950

Mean Density of 28,000 coliform per 100 ml (read on Y-scale).

THE SLOPE, $\sigma_{(Log)}$—The slope of the distribution line "A" determines $\sigma_{(Log)}$ which is given by scaling $\Delta Y'$ for a unit $\Delta X'$ in this example

$$\sigma_{(Log)} = .49$$

MEASURE OF STABILITY OR CHANGE IN MEAN DENSITY—The slope expected for a *3 portion* test procedure as transferred from the standard slopes of Fig. 18, is indicated on Fig. 20 as the dotted line "B". The actual distribution line "A" is steeper than that expected for a 3 portion test procedure which indicates at once that changes in level of Mean Density, and therefore real changes in quality, took place during the sampling period.

The question now arises: Within what range did the level of Mean Density shift? The answer to this depends upon a confidence we wish to have in the range defined. A range for a 90 per cent confidence would have a lower limit not exceeding 5 per cent and an upper limit not exceeding 95 per cent. Such a range in Mean Density is readily located graphically by projecting parallel to the standard slope line "B" from the intersections of 5 and 95 per cent with the distribution line "A". Probability of 5 per cent intersects line "A" at (a) (4,400); projecting parallel to line "B" to probability of 50 per cent locates the lower limit of Mean Density at (w) (14,500).

Similarly, probability 95 per cent intersects line "A" at (b) (175,000); projecting parallel to line "B" to probability 50 per cent locates the upper limit of Mean Density at (z) (52,500). This is to say that if 5 per cent of the MPN values do not exceed 4,400, (a), and the quality of water remained stable during sampling, the distribution line would fall along (a) (w) and would define a Mean Density of 14,500 located by (w).

Similarly if 95 per cent of the MPN values did not exceed 175,000, (b), and quality remained stable the distribution line would fall along (b) (z) and would define a Mean Density of 52,500 at (z). From this it is then reasoned that the probability is 5 per cent, that Mean Density was equal to or less than 14,500 and the probability is 5 per cent that Mean Density was equal to or greater than 52,500; or the probability is 90 per cent that Mean Density varied within the range 14,500 to 52,500, between (w) and (z).

In a similar manner the range for any confidence can be defined. We are 50 per cent confident that the Mean Density varied within the range 21,000 to 36,000; or 98

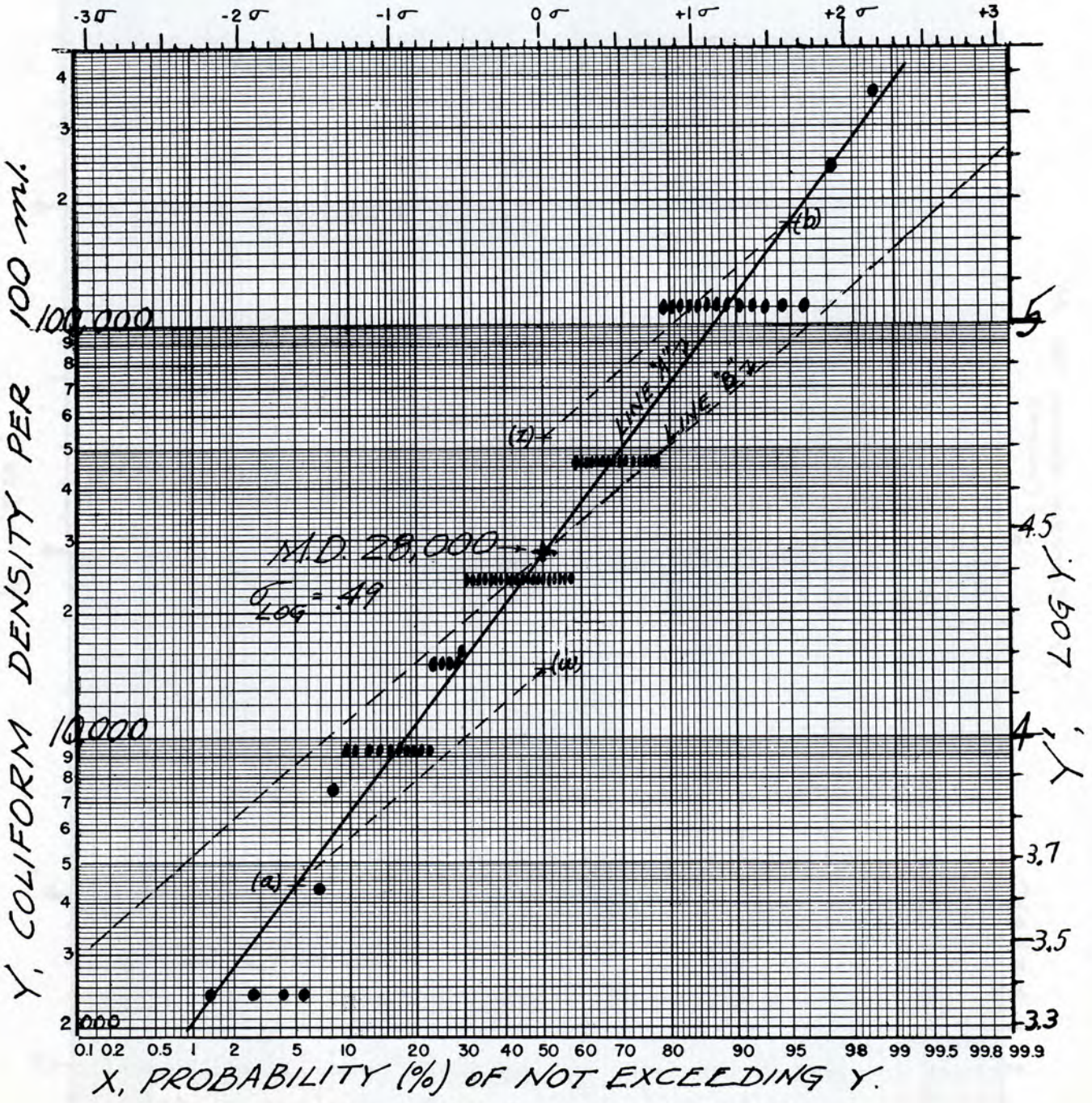


Fig. 20—Wyandotte, Mich.—New Intake—July 22 to Sept. 30, 1950

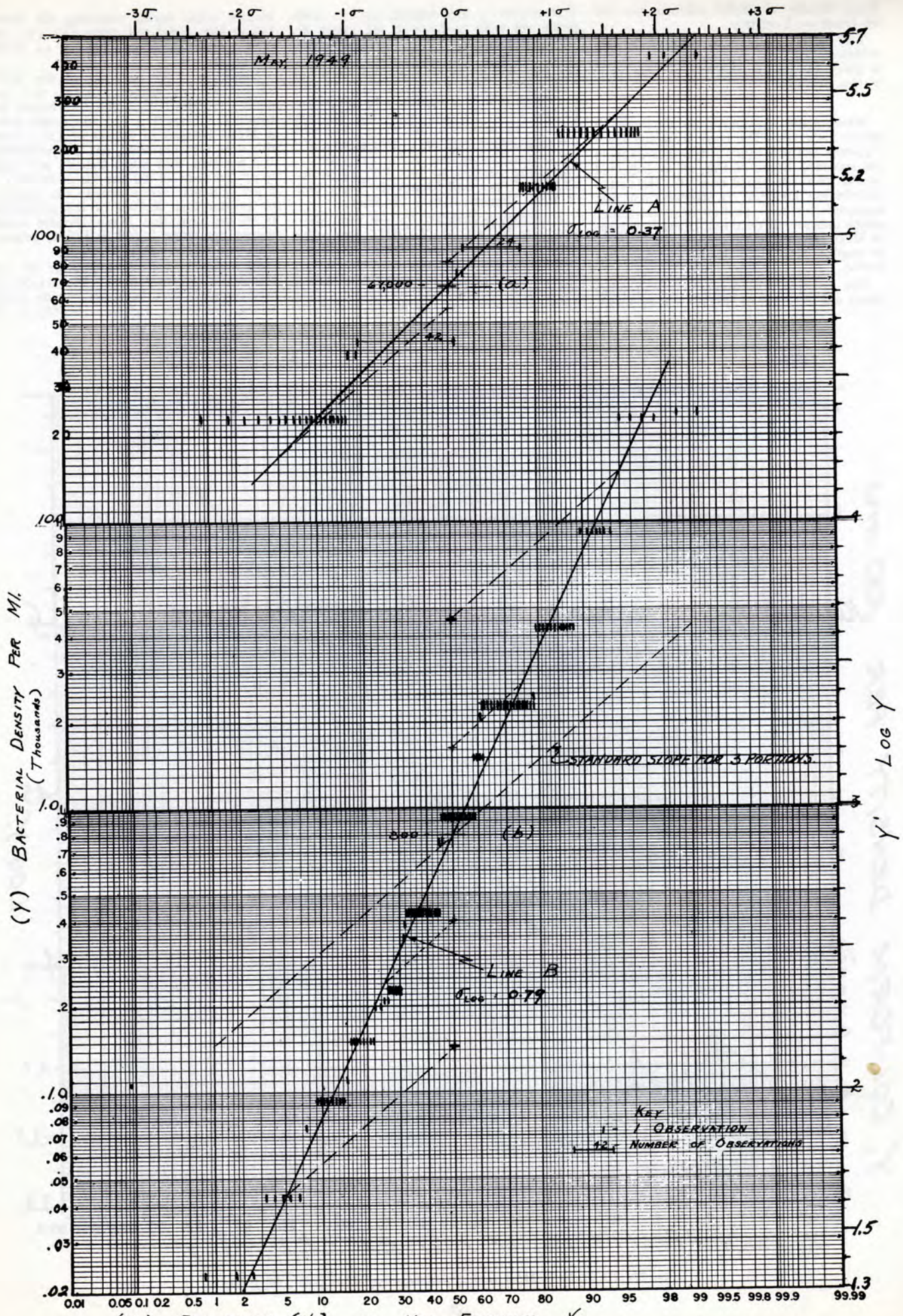


Fig. 21—Buffalo, N.Y., Sewage Plant Influent and Effluent Bacterial Density Distributions—May 1949

per cent confident that it varied within the range 11,000 to 70,000.

Since samples were taken at daily intervals, it can be reasoned that:

50 per cent of the days the Mean Density was above or below 28,000

50 per cent of the days the Mean Density was within the range 21,000 to 36,000

90 per cent of the days the Mean Density was within the range 14,500 to 52,500

98 per cent of the days the Mean Density was within the range 11,000 to 70,000

These ranges in Mean Density are a measure of the variation in real changes in level of quality, free of the influence associated purely with test procedure employed. It follows also that comparisons should be made in terms of Mean Density, not between individual MPN values.

Evaluation of Sewage Treatment Data

Fig. 21 illustrates an application of log-probability paper to the evaluation of coliform density of influent and effluent of Buffalo, N.Y.* sewage treatment works based upon 117 routine samples of 3 portions in each of 3 decimal dilutions taken 3 to 6 times daily except Sundays during the month of May 1949.

Line "A" is the distribution of the 117 MPN's of influent, indicating a reasonably straight line. The Mean Density is located at (a), 67,000 coliform per ml. (read on Y scale); and $\sigma_{(Log)}$ is 0.37.

The slope of line "A" is only slightly steeper than that expected for a 3 portion test procedure and hence Mean Density remained reasonably stable during the month, ranging from 56,000 to 80,000 for 90 per cent confidence.

Line "B" represents the distribution of 117 MPNs of effluent. Again a straight line is indicated with a Mean Density located at (b), 800 coliform per ml. (read on Y scale) and $\sigma_{(Log)}$ of 0.79. The slope of the effluent, line "B", is about twice as steep as that of the influent, line "A", and the standard slope expected for a 3 portion test procedure. This indicates at once that changes in level of Mean Density and therefore real changes in quality of effluent took place during the month. By projecting parallel to the standard slope for 3 portion test procedure from any desired probability intercept on line A to the 50 per cent probability gives the range in Mean Density for any desired confidence. If we assume that samples were taken at equal intervals (approximately true) it can be reasoned that

50 per cent of the time Mean Density was above or below 800 per ml.

50 per cent of the time Mean Density was within the range 400 to 1,600 per ml.

90 per cent of the time Mean Density was within the range 145 to 4,500 per ml.

The disinfection efficiency of the treatment system is determined by comparison of the Mean Densities of the influent and effluent, 68,000 per ml. and 800 per ml. respectively, indicating an average reduction in bacterial density of 98.8 per cent.

V—Tests for Statistical Significance

Comparison is a common method of appraisal. Observed differences or similarities may not be significant, as they may arise solely by chance. The purpose of this article is to present graphical methods to test the statistical significance of such apparent differences before drawing conclusions from comparisons.

Tests for statistical significance cannot *prove* a cause and effect relationship. The contribution of statistics is to define the expected variations due to chance. Actually, any difference in results, no matter how great, can always be ascribed to chance, but a point is reached where the probability that the difference is due to chance becomes so small that the investigator prefers to ascribe the difference to other causes. The decision that the difference is not due to chance and the subsequent search for the underlying *cause* of the difference is completely outside the realm of statistical technique and is entirely a matter of professional judgment of the specialist in the field. The statistical method is only a tool, not a substitute for reasoning.

*Data courtesy of George Fynn, Chief Chemist.

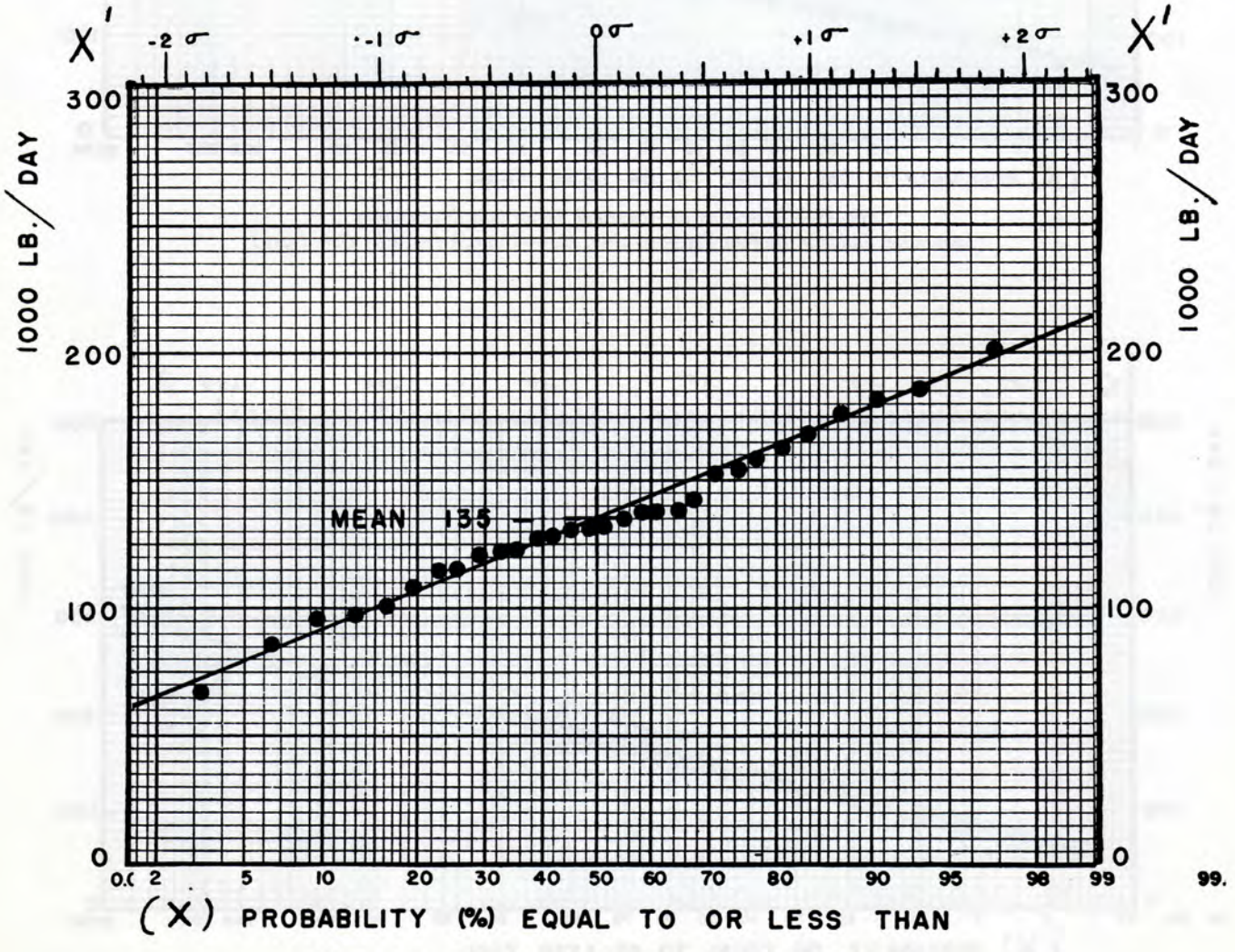


Fig. 22—Illustration of a Normal Distribution of Suspended Solids (Buffalo Sewage Works—Nov. 1939—Data Supplied by Geo. Fynn.)

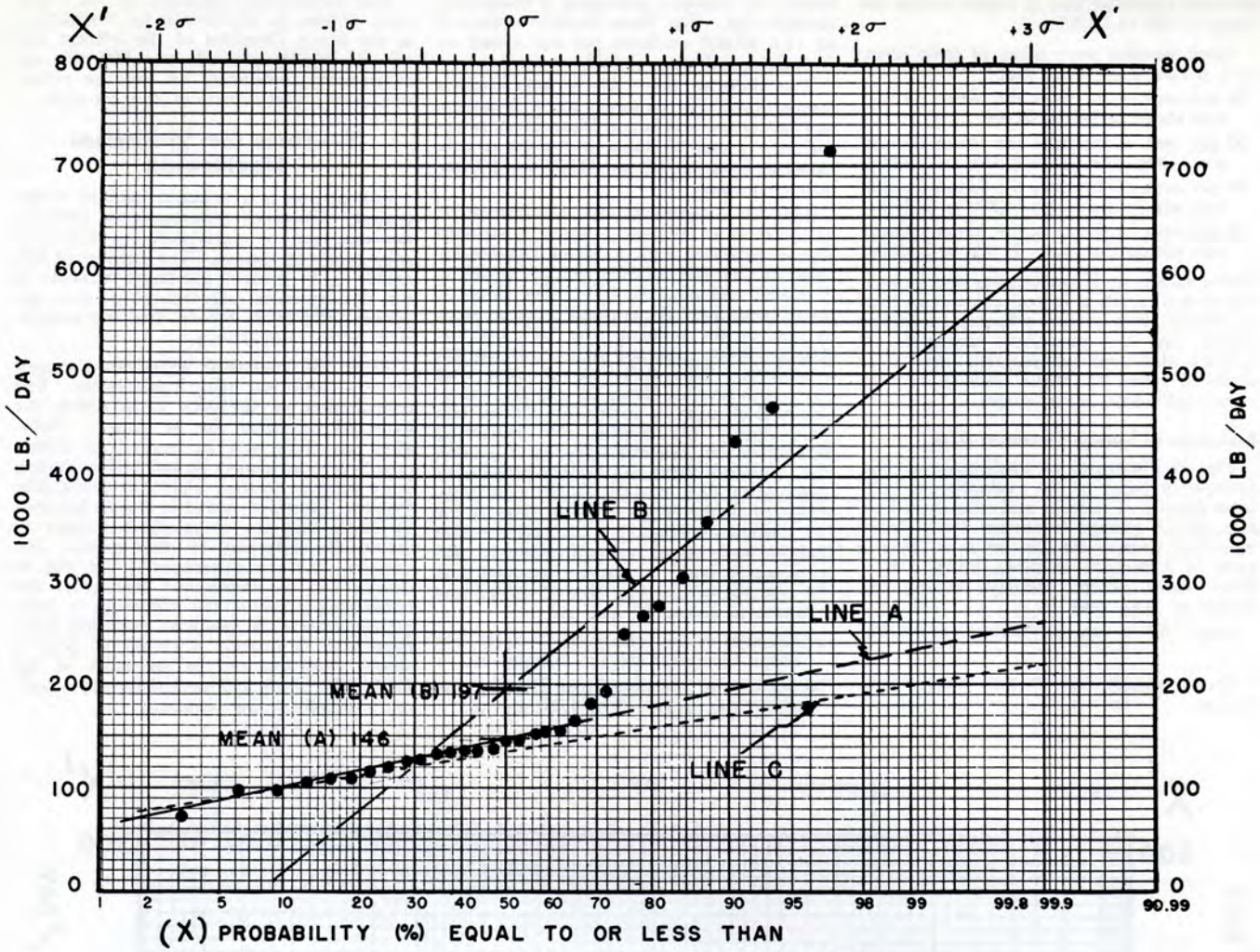


Fig. 23—Illustration of Abnormal Break in a Distribution
(Suspended Solids in Buffalo Sewage—Aug. 1939—Data Supplied by Geo. Fynn)

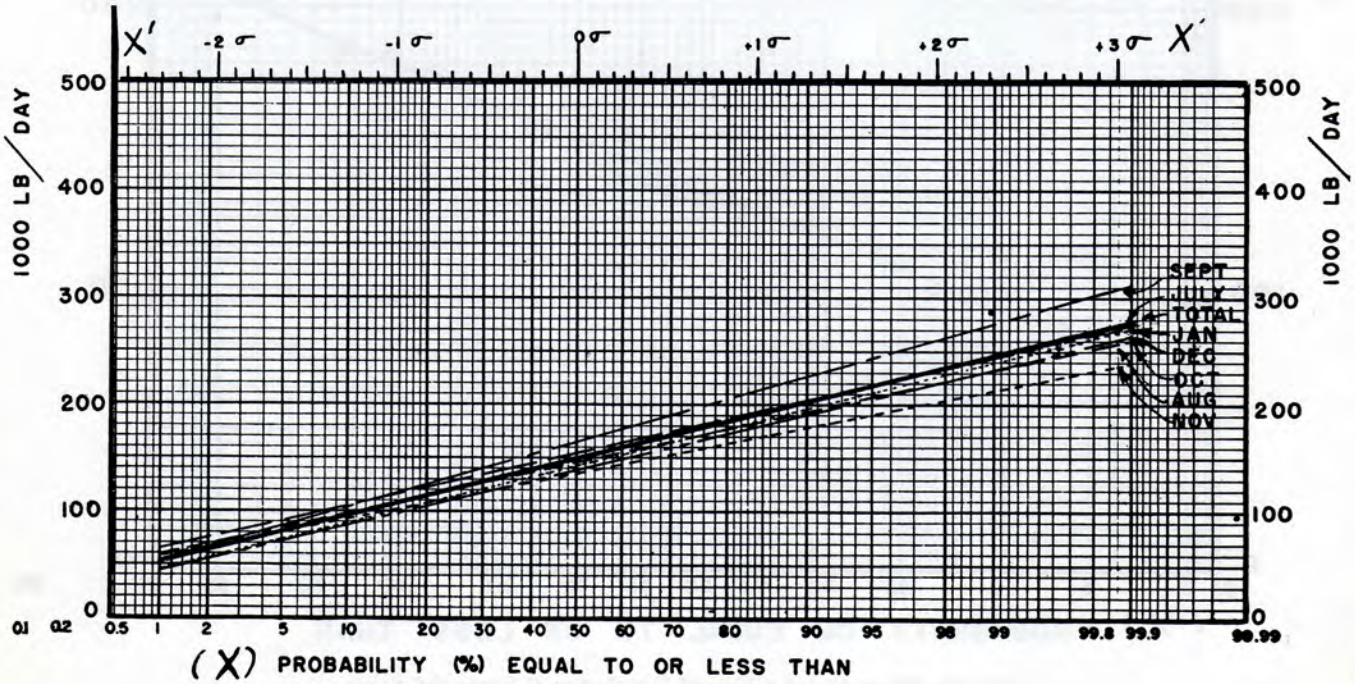


Fig. 24—Illustration of Graphical Test for Similarity Among Several Distributions

Testing a Single Series

Normality

The most important test applied to a single series is to determine if the distribution of individual measurements is *normal*. Graphically this is accomplished by plotting the data on probability paper described in the preceding articles. If a straight line is formed, the distribution is normal and symmetrical; if curved upward or downward, the distribution is skewed and asymmetrical. This is at once visually apparent. If a distribution plots as a straight line (is normal), then the raw data may be condensed, without loss of significance, to three summary figures:

(1) The Mean (Arithmetic Average)
(2) The Standard Deviation (σ)
(3) The Number of Observations (n)

Such a normal distribution is illustrated in Fig. 22, Suspended Solids for Buffalo Sewage for the Month of November 1939. Suspended solids are expressed on a quantitative basis computed from the concentration in parts per million and the daily volume of sewage in million gallons. Since a straight line is formed, the distribution is normal and may be summarized by the three statistics:

Mean ($\bar{y}$ graphical)	= 135.
Standard Deviation (σ graphical)	= 34.
Number of Observations (n)	= 30.

However, if a distribution does not form a straight line (is not normal), such summary statistics can be very misleading. The superiority of the graphical approach over the usual analytical methods is illustrated in the following examples.

Abnormal Breaks in Data

Routine water and sewage treatment works operating data are at times subject to abnormal breaks in otherwise normal variations. This is usually associated with sudden hydrologic changes which radically alter conditions. Figure 23, Line A, shows a distribution of daily suspended solids determinations for influent of Buffalo Sewage Treatment Works for the Month of August 1939. A sharp break from an otherwise normal distribution (Line A) is noted at the right upper end. Precipitation records for August show that the values which break away from the distribution were associated with the first day of rainfall flushing the combined sewer system.

The question now arises, how shall such data be summarized? The best summary is the graphical plot itself, as this is the only method clearly to differentiate between the normal and abnormal segments without loss of significance. Neglecting the graphical approach and attempting condensation of data by analytical statistical methods alone leads to two common misrepresentations:

1. Neglecting to test for normality and computing an arithmetic average and a standard deviation gives

Mean (Y computed)	= 197.6 Thd. lb. per day
Standard Deviation (σ computed)	= 137.3 Thd. lb. per day
Number of Observations (n)	= 31

Such an analytical summary would lead one to believe that the distribution was normal and, if this mean and standard deviation are reconstructed on Fig. 23, it would produce a distribution represented by Line B. Visually, it is apparent that Line B is too high and too steep and is obviously inconsistent with the bulk of the actual data. In attempting to define both the normal and the abnormal segments, it defines neither and, in fact, gives a false summary which is "nonsensical" in relation to the actual observed facts. Such a summary is dangerous.

2. Excluding the abnormal values and then computing an arithmetic average and a standard deviation gives

Mean ($\overline{Y}$ computed)	= 132.6 Thd. lb. per day
Standard Deviation (σ computed)	= 27.5 Thd. lb. per day

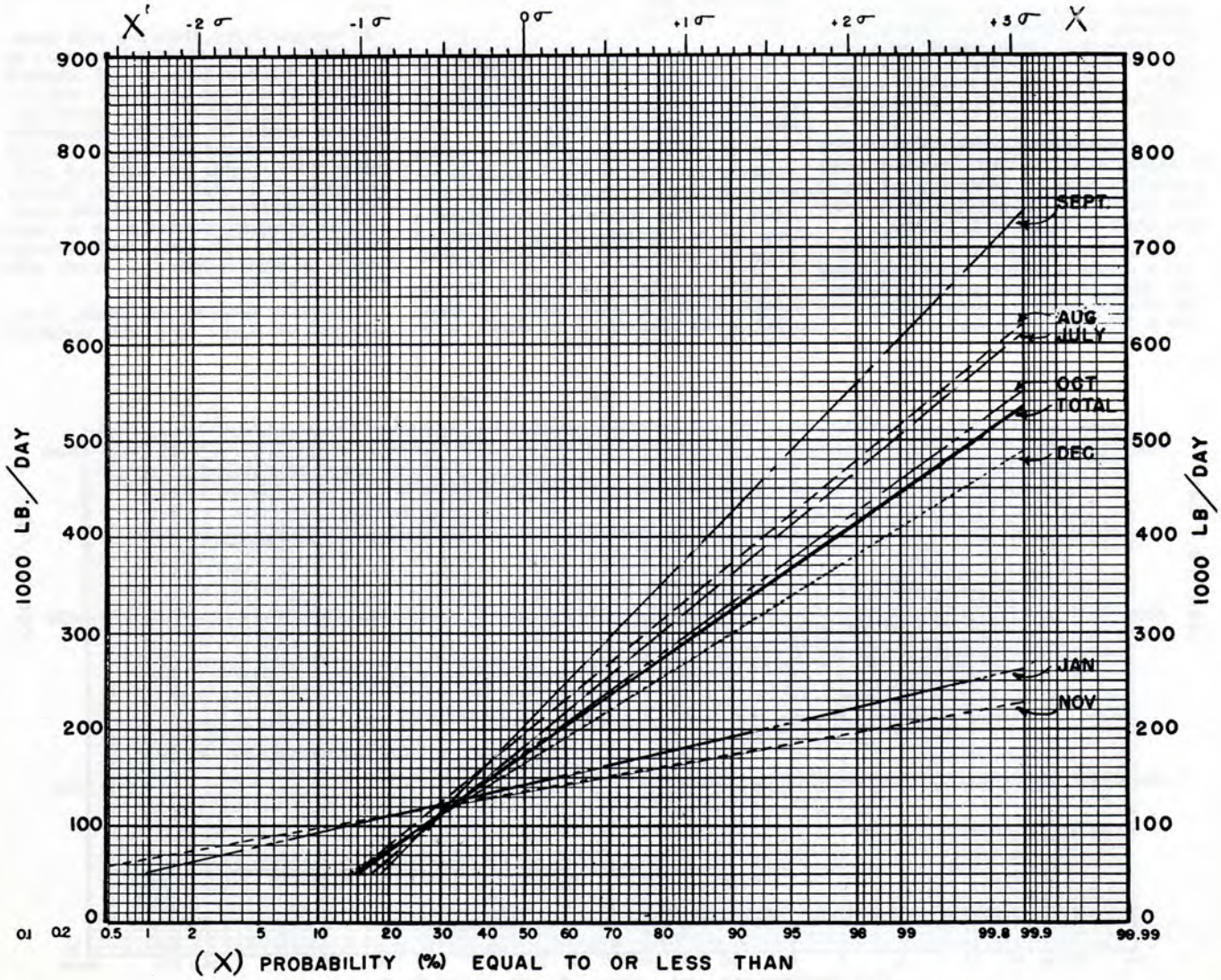


Fig. 25—Illustration of Graphical Test for Difference Among Several Distributions

Table 8

COMPARISON OF ANALYTICAL AND GRAPHICAL STATISTICAL SUMMARIES OF SUSPENDED SOLIDS

Buffalo Sewage Treatment Works Influent

July 1939-January 1940

Month 1939 (1)	Analytical Summaries: Computed mean 1000 lb./day (2)	Computed σ 1000 lb./day (3)	n (4)	Graphical Summaries: Graphical mean (Normal segment) 1000 lb./day (5)	Graphical σ (Normal segment) 1000 lb./day (6)	n (7)	Remarks (8)
July	182.9	139.3	31	153	39.0	31	85% normal followed by sharp break with 15% distinctly abnormal
August	197.6	137.3	31	146	37.3	31	65% normal followed by sharp break with 35% abnormal
September	206.1	173.5	30	166	48.0	30	75% normal followed by sharp break with 25% abnormal
October	174.0	123.1	31	139	40.0	31	80% normal followed by sharp break with 20% abnormal
November	134.9	30.8	30	135	34.	30	complete distribution normal
December	167.2	104.6	31	140	42.7	31	80% normal followed by sharp break with 20% abnormal
January, 1940	141.4	39.4	31	142	41.8	31	complete distribution normal
July, 1939-January, 1940	172.0	118.1	215	147	42.7	215	85% normal followed by sharp break with 15% abnormal

Number of observations (n) = 23. (8 largest values excluded)

This mean and standard deviation would reconstruct a distribution represented by Line C. Visually, Line C is too low and too flat. Excluding the 8 largest values completely ignores the abnormal segment and does not adequately represent the normal segment. Obviously, the deletion of the 8 largest values of a series of 31 values, even if they were in line with a normal distribution, would unduly reduce the mean and the standard deviation. "Throwing out" data is a dangerous procedure.

If a non-graphical condensation of data is insisted upon, it is best obtained through a graphical approach. Referring to Line A, Fig. 23, such a condensation of the raw data might be suggested as follows:

1. Approximately 65 per cent of the distribution is normal followed by a sharp break with the upper 35 per cent distinctly abnormal. (8 values distinctly out of line ranging from 248 to 717 Thd. lbs. per day)

2. The *normal* segment of the distribution is defined by

Mean ($\overline{Y}$ graphical) = 146 Thd. lb. per day

Standard Deviation (σ graphical) = 37.3 Thd. lb. per day

Total number of observations, normal and abnormal (n) = 31.

Such a summary, while no substitute for the actual published plot on probability paper, is at least consistent with the facts as observed and is not misleading.

As further emphasis of the value of the graphical approach, Table 8 contrasts the graphical summary statistics with the analytical summary statistics of the operating data for the Buffalo Sewage Treatment Plant for 7 months, July 1939-January 1940.

It will be noted from Table 8 that the normal suspended solids load to the treatment works from month to month, uninfluenced by rainfall flushes, is reasonably constant as reflected by the graphical means shown in Column (5). This is to be expected from a given population and normal industrial activity not subject to seasonal patterns. The day to day variability as reflected by the graphical standard deviation, Column (6), is also reasonably constant.

In contrast to this, there is a wide variation reflected in the analytical summary of computed monthly averages and standard deviations as shown in Column (2) and (3). November and January are of special significance because the complete distributions were normal without any breaks. Rainfall during these months was light and probably occurred as snow and hence flushing of the combined sewers did not take place. For the graphical summaries, it is noteworthy that the means and standard deviations of the other months tie in closely with these two months.

In contrast, however, the results of the analytical summaries show wide variations

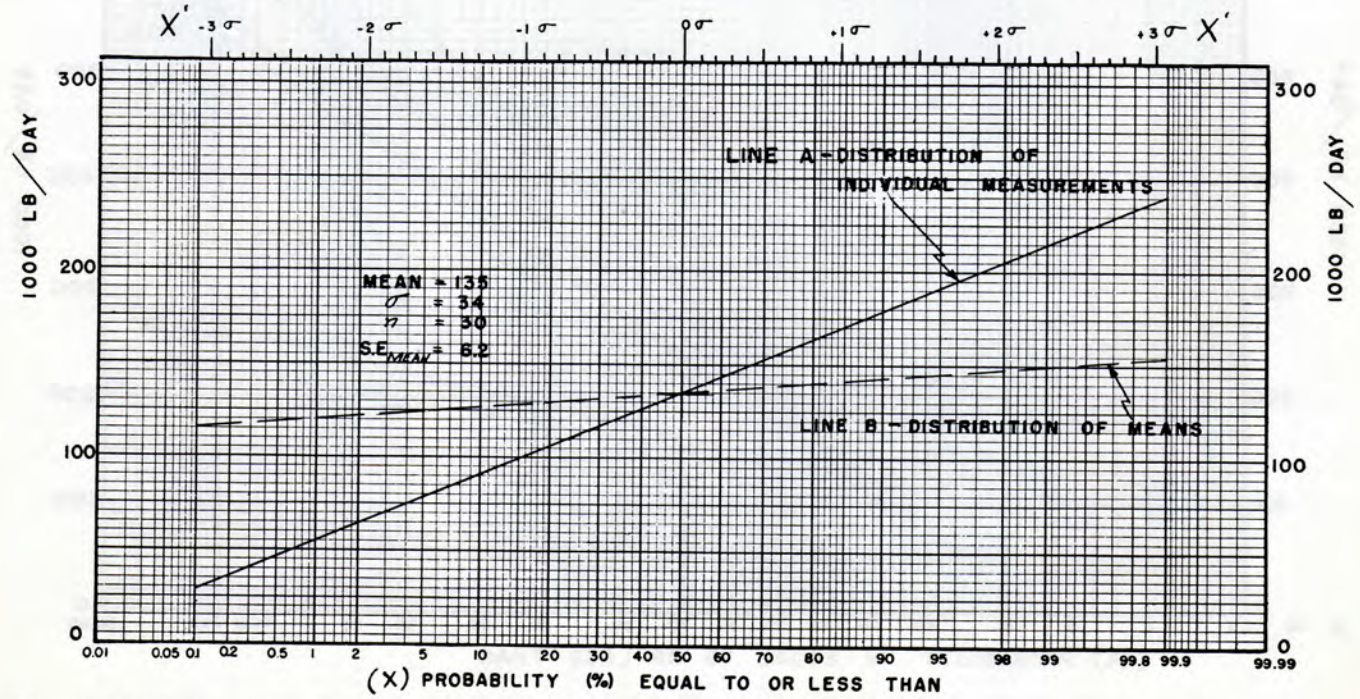


Fig. 26—Distribution of individual measurements and the distribution of means—suspended solids, Buffalo, N.Y. sewage, Nov. 1939.

in means and standard deviations for the other months when compared with November and January. Note, also, that only when the complete monthly distribution is normal (November and January) do the analytical results approach the graphical.

In appraising treatment works loading and performance, it is obvious that the all too frequently employed "Monthly Average" may be a very misleading figure. It is essential to differentiate between the normal and the abnormal before drawing conclusions or making comparisons.

Testing Two or More Series

Differentiating Among Series

Plotting two or more series of data on the same sheet of probability paper affords a quick visual appraisal of similarities or differences. For example, assume that the distributions summarized on Table 8 are all normal distributions as represented by the means and the standard deviations. Figure 24 represents the reconstruction of the normal distributions of the graphical summaries, Columns (5) and (6) in Table 8. The eight distributions cluster about the common mean and disclose similar slopes. The visual pattern strongly suggests similarity among the distributions.

In contrast, reconstructing distributions from the analytical means and standard deviations of Columns (2) and (3) in Table 8 (Fig. 25) indicates marked differences among the distributions, not only as to position of the mean but also as to slope. Such graphical differentiation affords a quick appraisal and points out areas where further testing between two distributions may be warranted.

Overlapping Test

Statistical significance of the difference between two series of data can readily be evaluated by the overlapping test. In comparing two series, the first operation is to differentiate them by plotting both on the same probability paper. However, it is not sufficient to demonstrate two distinct distribution lines to be confident that there is a statistically significant difference between the two series. We expect, by chance alone, that there will be shifts in the distribution lines upward and downward in plotting similar repeated sample series. Hence, to demonstrate a statistically significant difference, two distributions must remain beyond such shifts in position which chance alone can reasonably be expected to produce.

The Standard Error (S.E.) of the Mean is one measure of such shifts in a distribution line.

$$\underset{(\text{Mean})}{\text{S E}} = \frac{\sigma}{\sqrt{n}} \qquad (1)$$

The stability or reliability of the mean, as defined by its standard error, varies directly as the standard deviation (σ) of the individual measurements and inversely as the square root of the number of measurements in the series. The variability of the individual measurements is inherent in the phenomenon being observed and not much can be done about this. However, it is possible to control the reliability of a mean simply by increasing the number of individual measurements. Increasing the number of measurements four fold reduces the S E of the mean by ½; 16 fold by ¼; etc.; or more generally from the transformation of equation (1), the number of individual measurements required for a specified standard error of the mean is given by

$$n = \frac{\sigma^2}{(\text{S E})^2} \qquad (2)$$

For example (Referring to Fig. 26), the mean ($\overline{Y}$) and standard deviation (σ) as determined graphically are 135 and 34 Thd. lb. per day respectively for the series involving 30 individual measurements. From this single series it is then possible to evaluate the reliability of the mean

$$\underset{(\text{Mean})}{\text{S E}} = \frac{\sigma}{\sqrt{n}} ; \frac{34}{\sqrt{30}} = 6.2$$

Variation of the mean, like variation of the individual measurements, follows the normal probability distribution and from this it can be expected that for repeated series of 30 measurements

68% of the means would fall within the range of
135 ± 1 S E or 135 ± 6.2;
95% of the means would fall within the range of
135 ± 2 S E or 135 ± 12.4;
99.73% of the means would fall within the range of
135 ± 3 S E or 135 ± 18.6

This defines the range in shift upward or downward that can be expected by chance alone from plotting repeated series of only 30 individual measurements made under the same conditions.

The distribution of *means* can be represented graphically on probability paper sim-

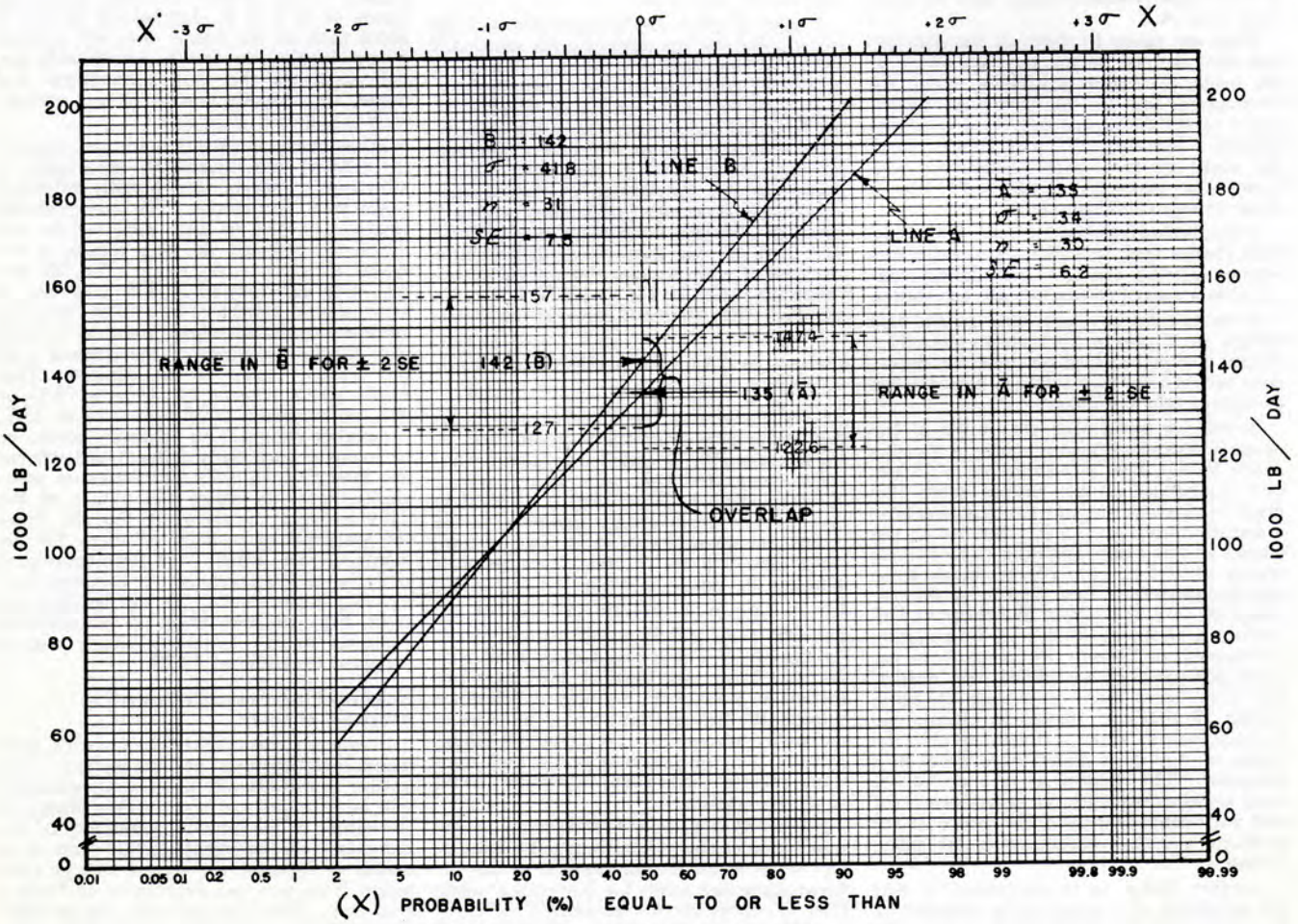


Fig. 27—Illustration of overlapping test between Nov. and Jan. distributions of suspended solids, Buffalo, N.Y. sewage.

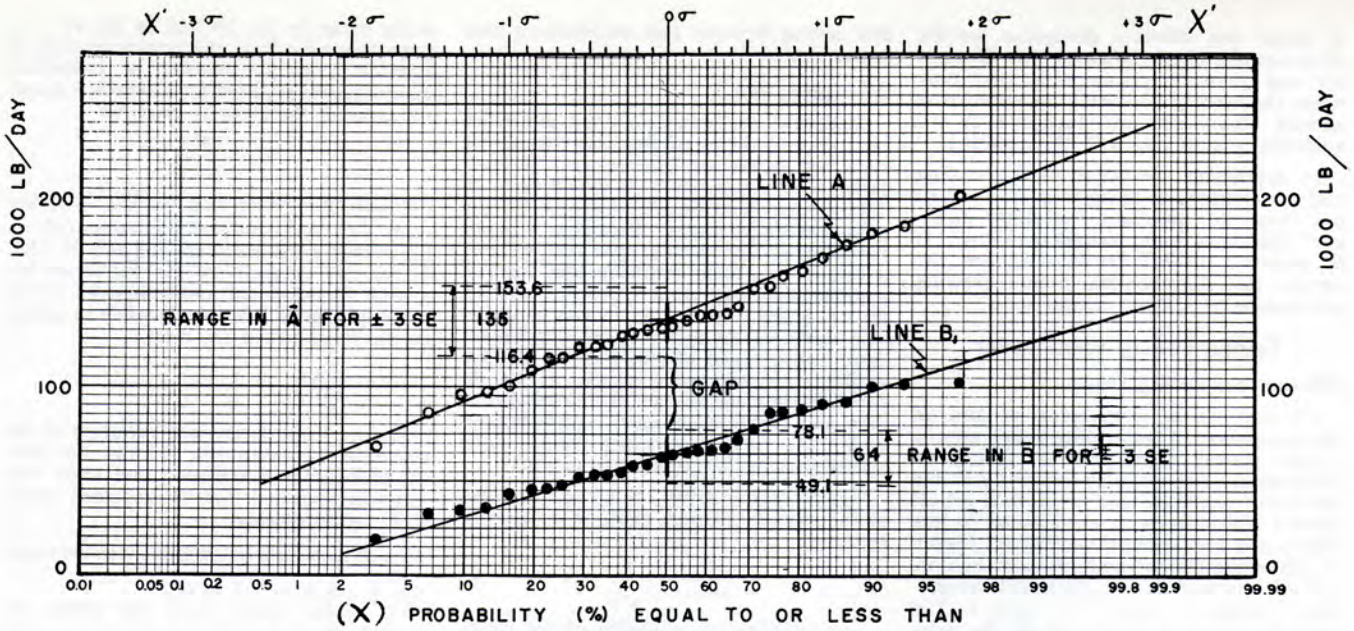


Fig. 28—Illustration of overlapping test between influent and effluent suspended solids, Buffalo, N. Y. sewage, Nov. 1939.

ply by laying off a line through the mean of a single series at a slope equivalent to the S E where X′ scale is in terms of S E in place of σ. The expected distribution of means from repeated series of 30 measurements is thus represented by Line B on Fig. 26. From this line, it is possible to read the range in means for any desired probability in the same manner that ranges in individual measurements can be read from Line A.

With the range in shifts of the distribution thus defined by the standard error of the mean, it is now possible to apply the overlapping test for the statistical significance of the difference between two distributions. Marking off these ranges about the mean of each series graphically discloses the variation expected by chance alone at any confidence level.

If no overlap develops between the ranges, then chance may be ruled out and the difference between the two series being compared is accepted as statistically significant.

If an overlap develops between the two ranges, or if one range encloses the other, chance cannot be ruled out, and the difference between the two series is not accepted as statistically significant.

It will be noted that the range of the mean is dependent upon a choice of a confidence level. The acceptance of a 95 per cent confidence range within which the mean is expected to vary by chance defines a specific range ± 2 S E. But ruling out chance at this range implies a risk of being wrong about 5 times in 100, as chance alone can be expected to show a mean outside the range of ± 2 S E about five times in 100. Setting a range of ± 3 S E is generally interpreted as "practically certain" to account for variation associated with chance, but even here, there is a risk of being wrong 27 times in 10,000, as chance can be expected to show a mean outside the range of ± 3 S E about three times in a thousand. The decision as to a confidence level always rests with the investigator and will vary depending upon the nature of the problem and the consequences of being wrong.

Another factor to be considered is that the reliability of a mean can be changed by the number of individual measurements. Hence, comparison between two means, each based upon a small number of observations, may develop an overlap and for these sized series, the difference would be considered not statistically significant. Increasing the number of measurements would reduce the range and may conceivably not develop an overlap and for the larger series, the difference then would be considered statistically significant.

Figure 27 illustrates an application of the overlapping test in comparing the suspended solids of the influent for the month of November with that of the month of January. Line A represents the distribution of the 30 individual measurements for November giving a mean of 135 and a standard deviation of 34 Thd. lb. per day. Line B represents the distribution of the 31 individual measurements for January giving a mean of 142 and a standard deviation of 41.8 Thd. lb. per day. Two distributions are formed with an apparent difference between the means of 142 — 135 or 7 Thd. lb. per day. The standard error of the means are $\frac{34}{\sqrt{30}}$ or 6.2 for November; and $\frac{41.8}{\sqrt{31}}$ or 7.5 for January. The question now is, is this difference statistically significant or can it reasonably be ascribed to chance variations associated with such small number of observations? In making a decision in this instance about chance variation in the means, we do not wish to be wrong more often than 5 times in 100. This requires defining the range in means by ± 2 S E; 135 ± (2 x 6.2) or 122.6 to 147.4 for $\bar{A}$; and 142 ± (2 x 7.5) or 127 to 157 for $\bar{B}$. Laying off ± 2 S E about each mean produces a large overlapping and therefore there is *no* statistically significant difference between the two distributions. Chance alone can easily account for the apparent difference, in fact the positions of the two lines could reverse themselves; November could be above January.

Figure 28 is a second illustration of the application of the overlapping test. Line A represents again the distribution of the influent suspended solids for November, while Line B represents the suspended solids of the effluent for November. The mean of the influent is 135 with a S E of 6.2 Thd. lb. per day, while the effluent mean is 64 with an S E of 4.7 Thd. lb. per day. As is to be expected, there is a substantial difference between the influent and the effluent distribution lines (135 — 64) or 71 Thd. lb. per day. In this instance, we wish to be "practically certain" (99.73%) that the difference is beyond what can be ascribed to chance. This defines a range about the means of ± 3 S E. Laying off ± 3 S E about each of the means does *not* produce an overlapping, in fact a considerable gap still separates the two distributions and hence, we conclude that there *is* a statistically significant difference.

Since in this illustration we are comparing influent and effluent, an opportunity is afforded to appraise the monthly efficiency of the treatment works. The most probable efficiency would be that reflected by the difference between the two means, a removal of (135 — 64) or 71 Thd. lb. per day or an efficiency of 71/135 or 52.5%. A very severe criterion by which monthly efficiency could be judged would be to compare the lowest range of the influent with the highest range of the effluent. This would give a removal of (116.4 — 78.1) or 38.3, an efficiency of 38.3 ÷ 116.4 or 33%.

Another approach to efficiency would be to compare each daily influent and effluent and then plot the daily efficiencies on probability paper to define the nature of the daily variation, as shown in Fig. 29. Here the graphical mean daily efficiency for the month is 53%, which is in close agreement with the 52.5% previously determined from monthly mean results. It will be observed from Fig. 29 that there is considerable variation in day to day efficiency with a standard deviation of 12.7%.

Number of Measurements Controlling Reliability of Mean

Frequently one is asked the question, how many measurements are necessary? The answer to this depends upon what reliability will be accepted and also foreknowledge of the nature of the inherent variability of the individual measurements. The first is a matter of decision, the second may be estimated from previous experience or from a trial run. Take, for example, the monthly mean influent suspended solids for Novem-

ber, 135 Thd. lb. per day with a standard deviation of 34. If it is decided to control the variation of the mean to a standard error of 2, how many individual measurements will be required? Accepting the trial σ of 34 and the specified S E of 2, the required number of individual measurements is then readily determined from equation (2)

$$n = \frac{\sigma^2}{(S\,E)^2} \text{ or } \frac{34^2}{2^2} \text{ or } 289.$$

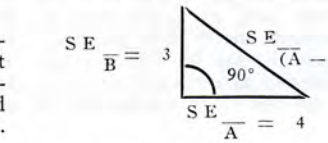
Or let us suppose it is decided to be "practically certain" that the mean would not vary more than ± 3 Thd. lb. per day. "Practically certain" is equivalent to ± 3 S E and hence 3 S E = 3, or the required S E is 1. For this specification

$$n = \frac{34^2}{1^2} = 1156$$

Difference Between Two Means

The simple overlapping test is a severe criterion for determining the statistical significance of a difference between two distributions. The assumption is made that the distributions approach each other; that is, as the smaller increases, the larger decreases. This requires a relative shift of 180° and hence the distance required to keep the distributions from overlapping is some increment of the sum of the Standard Errors of the two distributions. For example, if the Standard Error of the mean of Series A is 4 units and that of Series B is 3 units, 68 per cent certainty of a statistically significant difference would require the two distributions to be at least 3 + 4 units apart; 95 per cent certainty, 2 (3 + 4) apart; and "Practically Certainty" (99.73 per cent), 3 (3 + 4) apart. The overlapping test, therefore, defines the Standard Error of a *difference* between two means as the sum of Standard Errors of the two means.

If two series are free to shift independently of each other, they need not necessarily always vary by 180°. Actually the independent shifting of two distributions may conceivably have an average variation in relation to each other at 90°. The relative difference between two means then would be equivalent to the hypotenuse of the right triangle with the legs comprised of the two Standard Errors of the means.



Now the Standard Error of the difference between the two means is the square root of the sum of the squares of the two Standard Errors of the means, which is the hypotenuse distance.

$$S\,E_{\overline{A}-\overline{B}} = \sqrt{(S\,E_{\overline{A}})^2 + (S\,E_{\overline{B}})^2}$$

or in the example:

$$\sqrt{4^2 + 3^2} = \sqrt{25} = 5$$

It will be noted that while the Standard Error of the difference between the two means by the overlapping test is 3 + 4 or 7, by assuming complete independence the Standard Error of the difference is

$$\sqrt{3^2 + 4^2} \text{ or } 5.$$

Thus the overlapping test is unduly severe.

Assuming the independence between two means, employing the hypotenuse as the unit of Standard Error of the difference, the test for statistical significance between two means would proceed as follows:

1. Plot the two distributions to test for normality.
2. If normal, obtain the mean and standard deviation (σ) of each series.
3. Compute the Standard Error of each mean.

$$S\,E_{Mean} = \frac{\sigma}{\sqrt{n}}$$

4. Compute the Standard Error of the difference between the two means.

$$S\,E_{Diff.} = \sqrt{(S\,E_{Mean_1})^2 + (S\,E_{Mean_2})^2}$$

5. Obtain the difference between the two means.
6. Decide upon a confidence level for test criteria, 95 per cent certainty, or 99.73 per cent or any other level desired.
7. Apply the test for statistical significance of the difference.

For example, 95 per cent certainty:

Subtract from the difference, 2 times the Standard Error of the difference and, if a positive difference remains, it can be concluded that the difference is statistically significant at this level of confidence. If, after subtracting, no positive difference remains or if a negative value occurs, then the difference was completely washed out by chance alone and it can be concluded that the apparent difference between the two means is *not* statistically significant.

Applying this criterion to the two illustrations employed in the overlapping test: From Fig. 27

Series A (November)	Series B (January)
Mean = 135	Mean = 142
S E $_{\overline{A}}$ = 6.2	S E $_{\overline{B}}$ = 7.5

Difference = 142 − 135 = 7

$$S\,E_{(Difference)} = \sqrt{6.2^2 + 7.5^2} = 9.7$$

Test at 95% certainty:

$$7 - (2 \times 9.7) = -12.4$$

The difference is completely washed out by

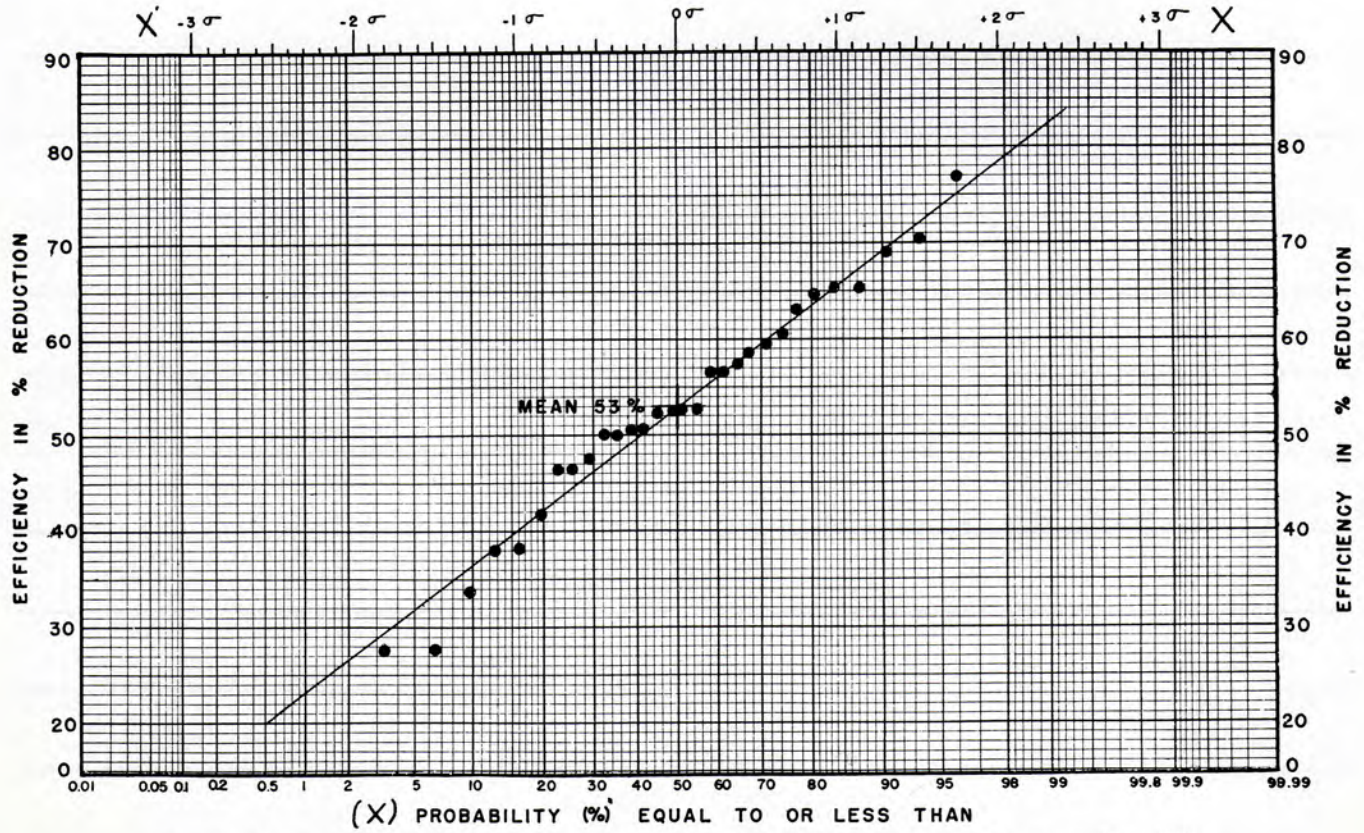


Fig. 29—Distribution of daily efficiency in removal of suspended solids—Buffalo, N.Y. sewage, Nov. 1939.

chance alone; therefore the apparent difference is *not* statistically significant. From Figure 28.

Series A Influent: Mean = 135, $S.E._{\bar{A}}$ = 6.2

Series B Effluent: Mean = 64, $S.E._{\bar{B}}$ = 4.7

$$\text{Difference} = 135 - 64 = 71$$

$$S.E._{(\text{Difference})} = \sqrt{6.2^2 + 4.7^2} = 7.8$$

Test at (99.73%) "Practical Certainty":

$$71 - (3 \times 7.8) = +47.6$$

A substantial positive difference remains after deducting practically all that can be accounted for by chance; therefore there *is* a statistically significant difference.

Historical

In concluding this series, reference to a few pioneering applications of statistics to sanitary engineering is appropriate. The profession is indebted to the early pioneering work of Professor Earle B. Phelps who made wide use of statistical tools; to the late Allen Hazen for his development of probability paper and its application to storage requirements and to flood probability; to Professor Gordon M. Fair for his early use of statistics as a research tool; to C. E. Keefer for his early application of probability paper to evaluation of sewage treatment operating data; and more recently to Professor E. J. Gumbel for his development of standard distribution of extreme values.

In preparation of this series, the writer wishes gratefully to acknowledge the assistance rendered by three Georges: i.e., Symons, (Man. Ed., *Water & Sewage Works*), Fynn (Ch. Chem., Buffalo Sewer Authority) and Hazey (Ch. Oper. Filter Plant, Wyandotte, Mich.) for supplying a considerable volume of data for analysis. The writer wishes also, gratefully to acknowledge the extensive and painstaking computations of Mrs. Mildred Harter.

NOTES