Proceedings of the International Congress of Mathematicians
Vancouver, 1974

Mathematics and the Picturing of Data*

John W. Tukey

1. Introduction. Why am I writing on this topic? Partly because picturing of data is important. Partly because, if present trends continue, an increasing fraction of all mathematicians will touch—or come close to touching—data during the next few decades. Mathematicians have many advantages in approaching data—and one major disadvantage. Those mathematicians who might come close to data need to know their advantages from their disadvantages.

Experience and facility with clear thinking—and with varied sorts of calculi that lead step-by-step from start to conclusion—knowledge of a variety of mathematical structures—even some of the more abstract are sometimes relevant to data—these are great advantages. The habit of building one technique on another—of assembling procedures like something made of erector-set parts—can be especially useful in dealing with data. So too is looking at the same thing in many ways or many things in the same way; an ability to generalize in profitable ways and a liking for a massive search for order. Mathematicians understand how subtle assumptions can make great differences and are used to trying to trace the paths by which this occurs. The mathematician's great disadvantage in approaching data is his—or her—attitude toward the words "hypothesis" and "hypotheses".

I must diverge for a moment to tell a story, dating to about 1946. The late Walter Mayer, then a member of the School of Mathematics at the Institute for Advanced Study, and I were chatting at the A.M.S. annual meeting at Rutgers. He was surprised that I was going to stay with Bell Laboratories, as well as with Princeton University. He explained how he had become involved with applied matters in Germany during World War I, and how happy he was to get back where,'and I

*Prepared in part in connection with research at Princeton University supported by the U.S. A.E.C.

quote, "If I say a g_{ik} has certain properties, it does". If you cannot occasionally modify the attitude Walter Mayer then expressed, work close to data may not be your forte.

When you come to deal with real data, formalized models for its behavior are not *hypotheses* in the mathematician's sense—in the sense that Walter Mayer so enjoyed—the language adopted by classical mathematical statistics *notwithstanding*. Instead these formalized models are *reference situations*—base points, if you like—things against which you compare the data you actually have to see how it differs. There are many challenges to all the skills of mathematicians—*except* implicit trust in hypotheses—in doing just this.

Since no model is to be believed in, no optimization for a single model can offer more than distant guidance. What is needed, and is never more than approximately at hand, is guidance about what to do in a sequence of ever more realistic situations. The analyst of data is lucky if he has some insight into a few terms of this sequence, particularly those not yet mathematized.

2. Picturing in simple cases. Picturing of data is the extreme case. Why do we use pictures? Most crucially to see behavior we had not explicitly anticipated as possible—for what pictures are best at is revealing the unanticipated; crucially, often as a way of making it easier to perceive and understand things that would otherwise be painfully complex. These are the important uses of pictures.

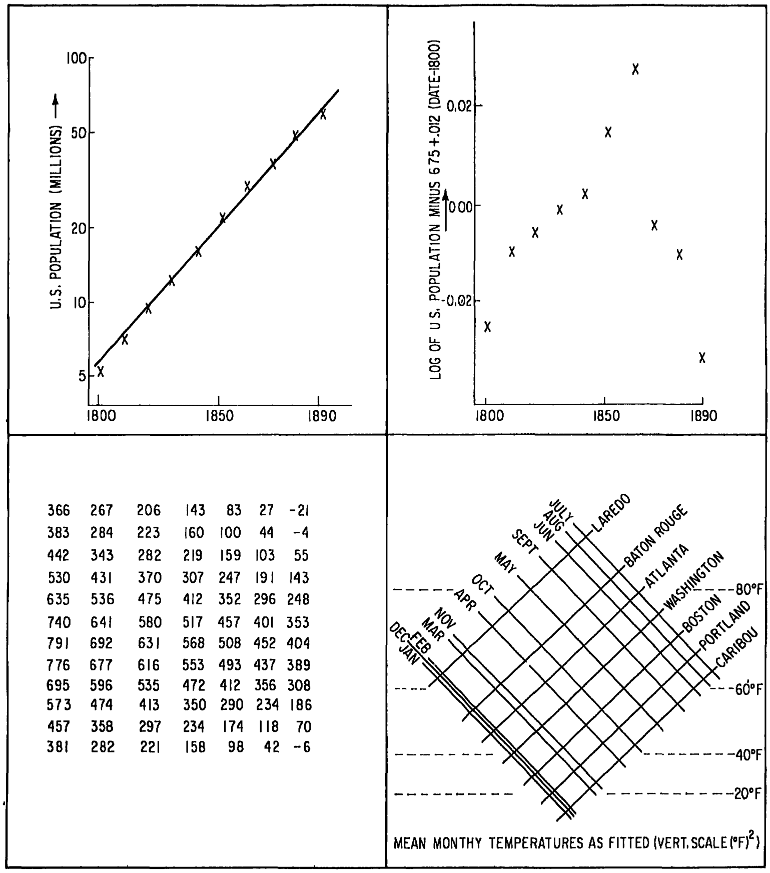
We can, and too often do, use picturing unimportantly, often wastefully, as a way of supporting the feeble in heart in their belief that something we have just found is really true. For this last purpose, when and if important, we usually need to look at a *summary.*

Sometimes we can summarize the data neatly with a few numbers, as when we report:

a fitted line—two numbers,
an estimated spread of the residuals around the "best" line—one more number,
a confidence interval for the slope of the "best" line—two final numbers.

When we can summarize matters this simply in numbers, we hardly need a picture and often lose by going to it. When the simplest useful summary involves many more numbers, a picture can be very helpful. To meet our major commitment of asking what lies beyond, in the example asking "*What is happening beyond what the line describes*?", a picture can be essential.

The NW corner of Figure 1 is a wasteful, unhelpful picture, except for those who must be reassured that the U.S. population increased, roughly exponentially, between 1800 and 1890. The NE corner is an effective, helpful picture, showing how the census population deviated from one exponential. ("Census population," because census errors are comparable with what we now see.) It shows us the comparison between the data and a simple, well-understood reference. The SW corner gives, numerically, the simplest relatively close fit to some data. Who among us can look at this and tell what is going on? The SE corner shows graphically exactly the same fit. All we have to do to use it is to learn to pay no attention to horizontal



366	267	206	143	83	27	-21
383	284	223	160	100	44	-4
442	343	282	219	159	103	55
530	431	370	307	247	191	143
635	536	475	412	352	296	248
740	641	580	517	457	401	353
791	692	631	568	508	452	404
776	677	616	553	493	437	389
695	596	535	472	412	356	308
573	474	413	350	290	234	186
457	358	297	234	174	118	70
381	282	221	158	98	42	-6

FIGURE 1

position, to forget that coordinate entirely. Such a picture makes the otherwise complicated understandable.

The main tasks of pictures are then:

to reveal the unexpected,

to make the complex easier to perceive.

Either may be effective for that which is important above all: *suggesting the next step in analysis, or offering the next insight.* In doing either of these there is much room for mathematics and novelty.

How do we decide what is a "picture" and what is not? The more we feel that we can "taste, touch, and handle" the more we are dealing with a picture. Whether it looks like a graph, or is a list of a few numbers is not important. Tangibility is important—what we strive for most.

The great geologist Chamberlain once said, in a paper recently reprinted in Science after seven decades: "Science is the holding of multiple working hypotheses". We need to go further, to the broader—prouder—maxim: "The picturing of data must be sensitive, not only to the multiple hypotheses we hold, but to the many more we have not yet thought of, regard as unlikely or think impossible".

3. Some details. Just using residuals as a way of looking at the data against the straight line—instead of through its appearance—was not difficult. The same idea recurs, however, with ever-increasing complexity. As mathematicians we are used to taking a bit of process from one procedure and putting it in others. We need to do this widely and subtly with

$$\text{given} = \text{fit PLUS residual.}$$

Notice that we have pictured "plus" by writing it out in capital letters—thus giving it its proper emphasis instead of using a single unobtrusive symbol.

Consider a fit in the form row PLUS column. (By not writing $a_i + b_j$ I am picturing the formula—and in the process not leaving its essentials to subscripts and + signs which the nonmathematician finds it easy not to notice.) When we plot the points with coordinates (row, column) (easily found as the intersection of a family of vertical lines with a family of horizontal lines), the loci row + column = constant are parallel straight lines with slope -1, and we have only to turn our picture through 45 degrees. All very simple—but equally mathematical.

We have now made a picture where the vertical coordinate is all meaningful and the horizontal coordinate *is to be forgotten.* We have had many years' experience with graphs where both coordinates have meaning. Many find it suprisingly hard to give up the idea that ALL quantitative pictures have to involve TWO meaningful coordinates.

4. Mental picturing of matrices. The mental pictures of those concerned with data have, of necessity, to be more or less mathematical. It is important that they involve appropriate mathematics, sufficiently understood. Failure to do this is most evident in connection with finite-dimensional linear spaces.

To be useful in dealing with data, understanding of matrices needs to be both abstract and concrete, and the bridge between needs to be well trodden in both directions. An abstract matrix—whether it represents a linear transformation or a quadratic form—need not involve a coordinate system, though it must involve two vector spaces (which may coincide). A numerical matrix—whether it represents a linear transformation or a quadratic form or a change of coordinates, three interpretations that we MUST keep clearly separate, must involve not only two specific vector spaces but two specific coordinate systems (which again may coincide). To hint by saying that a matrix is "$p \times p$", that any two matrices with matching numbers of columns and rows can be freely multiplied together, is an egregious source of false understandings and error.

The linear spaces that arise in treating data are rarely finite-dimensional examples of the familiar self-conjugate Hilbert spaces. Rather they are finite-dimensional Banach spaces. (1, 3, 5) is quite different as the values of (x_1, x_2, x_3) in some given data set than it is as the coefficients $\{c_i\}$ in $c_1x_1 + c_2x_2 + c_3x_3$. Inner products between a vector of c's and a vector of x's are well defined, and do not need assumptions. (Quadratic norms seem always to come from precise assumptions about relative sizes of variances and covariances. If these really mattered crucially, we would almost always be in very deep trouble. Fortunately, only rough ratios matter.)

Let me remind you of the distinction between orthogonality $(x_i, x_j) = 0$, for $i \neq j$ and biorthogonality $(c_i, x_j) = 0$, for $i \neq j$.

Many have heard of orthogonality; some bow down to it as to an idol. Fewer have heard of biorthogonality, yet it is a much more important idea in handling data. The practice of solving linear equations usually involves orthogonalization, yet who teaches that this is an arithmetically convenient route to biorthogonality? (At times, I suspect this is orthogonality's chief—if not only—virtue.)

Somehow ease of writing—or concentration on linear transformations—more precisely on linear transformations from and to the same space—rather than on quadratic forms—has made many find $|A - \lambda I| = 0$ the natural equation for eigenvalues and eigenvectors—which are then said to be "of A". When we deal with data, we are much more often concerned with $|B - \lambda C| = 0$ and we usually need to be concerned with the result as being "of B compared with C".

Some may think of these as small points. They are not tremendous, but their neglect or misinterpretation has kept many from an adequate understanding of what they are doing.

Mathematicians should, I believe, see that their students:

understand that simple matrix operations, like inversions, finding eigenvalues, or finding eigenvectors really exist, and can be used—by a computer-and-programs system when available, by hand in case of need,

understand the simple abstract characteristics of linear spaces, their conjugate spaces, linear transformations, and quadratic forms,

understand what the effect of change of coordinates is on such objects,

understand how to tie the concrete numerical and the abstract algebraic together, when thinking about and working with real examples,

understand that every numerical matrix implies two coordinate systems.

If people are to throw letters for matrices around as freely as they do letters for numbers—a noble goal—they need to be equally willing to throw in numerical values in the two cases—and as willing to introduce matrix coordinate changes in the one as to introduce scalar coordinate changes, perhaps feet to inches, in the other. Matrix algebra needs to have the same reality as scalar algebra.

5. Cumulations. How do we present information about distribution—whether the distribution is a mathematical entity, a Lebesgue measure, or a batch of points? By saying "information about" we are admitting the need to give only a summary. By imbedding this question in a paper whose title includes the word "data" we are admitting that such questions as whether the mathematical distribution is discrete, singular or absolutely continuous are not to be answered.

Before we summarize, the mathematician is likely to want to use the cumulative

(cumulative distribution function), usually defined with a $\leqq$. Both he and the practitioner can gain by the redefinition

$$F(x) = \text{Probability } \{y < x\} + \tfrac{1}{2}\,\text{Probability } \{y = x\}$$

which makes the cumulative of $-y$ exactly $1 - F(x)$, any discontinuities included, and makes Fourier inversion exact at any discontinuities. (Karl Pearson did this without saying so seven decades ago.)

The empirical distribution function is often defined as *1/nth of the Count of y's* $\leqq$ x where n is the total count. Again it is better to choose

$$n \cdot F_n(x) = (\text{Count of } y\text{'s} < x) + \tfrac{1}{2}\,(\text{Count of } y\text{'s} = x).$$

We will soon learn to change n to $n + \frac{1}{3}$ and add $+ \frac{1}{6}$ on the right.

How are we to summarize a function? This question is easier for the empirical case, where we have at most n values to begin with, where it is easy to move back from a higher-level description by a function to a lower-level description by n numbers, much more tangible as long as n is not too large.

If we draw a sample $y_1, y_2, \cdots, y_n$, of n observations from $dF(y)$ we are usually only concerned with symmetric functions of the y's. (Sometimes I think that only statisticians work with nonpolynomial symmetric functions.) The numerically least of $y_1, y_2, \cdots, y_n$ which we will label $y_{1|n}$ is surely a symmetric function of the y's. (Rearrangement surely makes no difference to its value.) The same is true of $y_{2|n}$, the next-to-smallest, and so on. The order statistics $y_{1|n} \leqq y_{2|n} \leqq \cdots \leqq y_{n|n}$ obtained by rearranging the y's in increasing order are actually the most general symmetric functions of $y_1, y_2, \cdots, y_n$.

Assume now that $y_i, \cdots, y_n$ are a random sample from a distribution with a continuous cumulative $F(x)$. This is a reference situation, not a real one, but that should not interfere with using it for guidance.

It is easy to see that the distribution of $F(y_{i|n})$ depends only on n and i, not at all on F. There are many ways to typify the location of each of these distributions with a number. The one that is most useful is by its median, especially since the operation of taking medians commutes with monotone re-expressions. Using the available tables—either of the beta distribution or of Snedecor's F distribution—we can learn that

$$\text{median } \{F(y_{i|n})\} \lessapprox \frac{i - \frac{1}{3}}{n + \frac{1}{3}}$$

where "$<$" means "is less than" and "$\approx$" means "is close to". Indeed for $n \geqq 5$ and all i, the reversed inequality holds when $(i - 0.3)/(n + 0.4)$ replaces $(i - \frac{1}{3})/(n + \frac{1}{3})$.

By the commutativity noted above, $F(\text{median}\{y_{i|n}\})$ satisfies the same condition, so that

$$\text{median } \{y_{i|n}\} \lessapprox F^{-1}\left[\frac{i - \frac{1}{3}}{n + \frac{1}{3}}\right].$$

Thus the natural fraction to associate with the ith of n is neither of the traditional choices—neither $(i - \frac{1}{2})/n$ nor $i/(n + 1)$—but rather $(i - \frac{1}{3})/(n + \frac{1}{3})$.

What does this tell us about empirical cumulatives? Almost certainly that, if we must use them,

$$F_n(x) = \frac{(\text{Count of } y\text{'s} < x) + \frac{1}{2}(\text{Count of } y\text{'s} = x) + \frac{1}{6}}{(\text{Total count of } y\text{'s}) + \frac{1}{3}}$$

is a still better choice. At the ith order statistic, the numerator is $(i - 1) + (1/2) + (1/6) = i - (1/3)$. So much for getting our ideas and definitions a little clearer.

6. Summarizing. The n order statistics are a natural beginning for summarization. We want to pick out some of them to stand for the rest. We can do this fairly well, because we know roughly what their correlation structure is: roughly equal correlations for equal values of i_1/i_2.

If we start with $i = 1$, and double, getting, successively, $i = 2, 4, 8, 16,\cdots$ or go up by half-octaves, getting, after rounding to integers, $i = 1$, blank, $2, 3, 4, 6, 8, 11,\cdots$ we will have sequences of roughly equally correlated order statistics. We can use such sequences, from above and from below, quite effectively as summaries.

This works well for some purposes. For others, where we would like to make use of the approximate stability of $(i - \frac{1}{3})/(n + \frac{1}{3})$ as n changes, we do better to begin in the middle and make our way toward both ends using:

depth of $y_{i|n}$ = the lesser of i or $n + 1 - i$ = result of counting in from the nearest end,

depth of median $= \frac{1}{2}$ (1 + total count),

depth of hinge $= \frac{1}{2}$ (1 + depth* of median),

depth of eighth $= \frac{1}{2}$ (1 + depth* of hinge),

$\cdots$

where we interpret depth* as "the integer part of the depth of" whenever, as is usual, we wish to confine i to integers or half-integers. (Otherwise we might solve $(i - \frac{1}{3})/(n + \frac{1}{3}) = 2^{-j}$ getting $i = (1/3) + (2^{-j}/3) + n/2^j$, to which the above is a reasonable approximation and which is more delicately precise than data warrants.)

As a result,.we summarize our distribution information with a sequence of about $2 \log_2 n$ values (about $2 \log_2 n + 2$ if we go all the way, including depths $1h$ and 1).

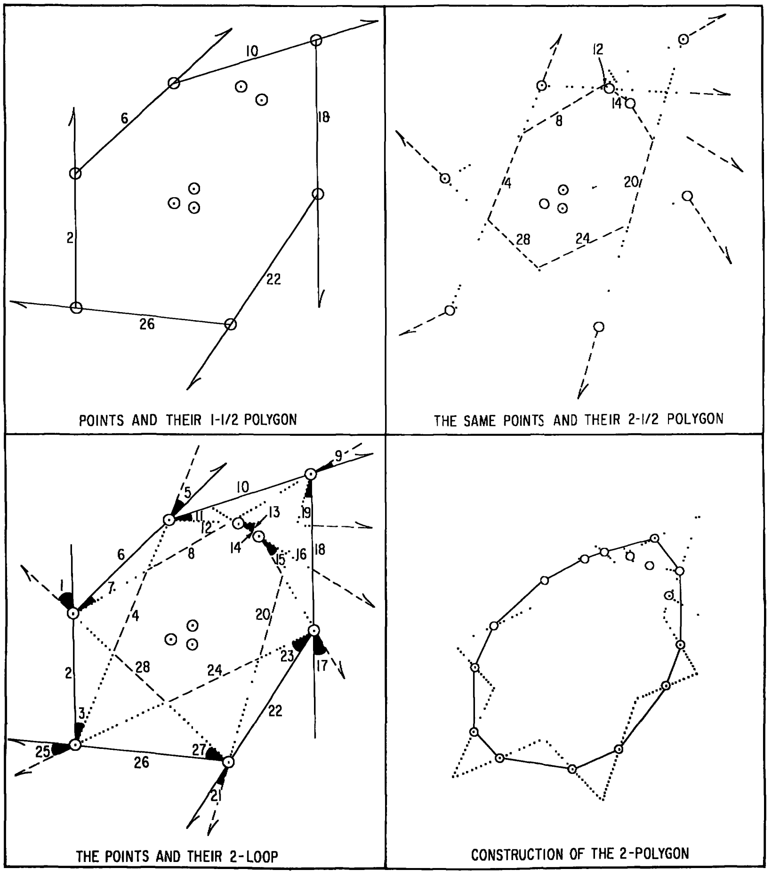
We have come a long way toward tangibility in going from $F_n(x)$ to $2 \log_2 n$ selected order statistics but this is only part way. We have not yet brought in a reference situation, so we are not in the favorable position of explicitly comparing what we have with a standard.

One approach to bringing in the Gaussian reference will be noted here, another later. If, for mathematicians' use only, we label the selected order statistics $L_k,\cdots,$ $L_3, L_2, M, U_2, U_3,\cdots, U_k$ where L_j and U_j share a depth, we can replace each pair by the corresponding "mid" and "spread"

$$M_j = (L_j + U_j)/2 \quad \text{and} \quad S_j = U_j - L_j$$

and if we then divide the spreads by the values of the corresponding spreads for the unit Gaussian distribution, we obtain two sequences of numbers such that the median values would be μ for each mid and σ for each divided spread *provided* we were sampling from a Gaussian distribution with mean μ and variance σ^2. Further investigation then suggests that we plot both sequences against the square of the divisor used for the corresponding spread. Thus, when we want to separate the issues of skewness from those of tail elongation, rather than separating what happens in the lower tail from what happens in the upper tail, we can reach a respectable pair of pictures comparing any batch of n observations with a Gaussian standard. (We can convert very easily to a logistic standard, or a Cauchy standard, or the standard provided by any other symmetric distribution. If a seriously asymmetric distribution is a natural standard, the separation into mids and spreads is likely to be unnatural.)

7. And what of two dimensions? How can we generalize all this to the plane? Specifically to the affine plane? Direct generalizations of order statistics fail miserably. But if we regard an order statistic $y_{i|n}$ as an oriented (up or down across the real line) point, with $\geqq i$ of the n values to its "left" or on it and $\leqq i - 1$ strictly



POINTS AND THEIR 1-1/2 POLYGON

THE SAME POINTS AND THEIR 2-1/2 POLYGON

THE POINTS AND THEIR 2-LOOP

CONSTRUCTION OF THE 2-POLYGON

FIGURE 2

to its "left", we can generalize easily. In the plane, an (i, j) line will be any directed line with $\geqq i$ points to its left or on it and $\leqq j$ points strictly to its left. For any (i, j) the set of (i, j) lines is closed. For any $i < n$, there is one and only one $(i, i - 1)$ line in a given direction, one line of depth i. Thus the $(i, i - 1)$ lines form a closed curve of lines of depth i, the i-loop. If $j < i - 1$, the set of (i, j) lines is finite. The set of $(i, i - 2)$ lines forms a closed polygon, the $(i - \frac{1}{2})$-polygon (all its sides belong to both the $(i - 1)$-loop and the i-loop.

The NW corner of Figure 2 shows 11 points and the $1\frac{1}{2}$ polygon; the NE corner the same points and the $2\frac{1}{2}$ polygon; the SW corner the two polygons and the sectors defining the segments of pencils of lines that complete the 2-loop.

(Each filled-in sector represents all directed lines through the vertex which pass out through the sector. If a segment of a pencil is a "stub", is each such loop a "polystub"?) Clearly the 2-loop is too complex to be a satisfactory generalization of a pair of order statistics of matched depth. The midpoints of the segments cut off by the $(i - \frac{1}{2})$-polygon from the extensions of the sides of the $(i + \frac{1}{2})$-polygon define a new, intermediate polygon, the i-polygon, which seems to be a satisfactory generalization. The SE corner of Figure 2 shows the 2-polygon for the 11-point example.

8. Probability plots. A naive answer to "How should we picture a distribution based on n values?"—essentially the question asked in §6—would be to say "make a probability plot". (There are many weaknesses of such plots.)

Today, Luis Nanni and I are hard at work developing a plot that meets these difficulties as well as we can see how to do this. Why did it take so many decades to attack this problem? Presumably because too few of us have tried to make more useful pictures.

PRINCETON UNIVERSITY
PRINCETON, NEW JERSEY 08540, U.S.A.

BELL LABORATORIES
MURRAY HILL, NEW JERSEY 07974, U.S.A.