8

Data Analysis and Behavioral Science or Learning to Bear the Quantitative Man's Burden by Shunning Badmandments

J. W. Tukey

PREFACE

The basic attitudes of this account were formed by nearly twenty years of experience with varied applications of statistics to the analysis of data. The core of insight around which these attitudes developed came from working alongside Charles P. Winsor during World War II. Their rather uninhibited expression was aided by a year's experience (1957-58) as a Fellow of the Center for Advanced Study in the Behavioral Sciences, which led to four recognition, viz: A recognition that the failures of behavioral scientists to get the most out of statistics were essentially similar to the failures in other fields. A recognition that very able behavioral scientists lacked certain insights which are conveyed by example and osmosis (and not by precept or word) to many physical

The Collected Works of John W. Tukey (1986) Wadsworth, Inc., Belmont, CA.

Previously unpublished manuscript (1961).

scientists, even to those of rather limited ability. A recognition that, in the areas where these insights are conveyed implicitly, mathematical formulas are often more important carriers than words. A recognition that there are many scientists (not all in behavioral science) for whom words are essential if the message is to reach them, and that almost everyone is helped a little when it is expressed in words. This year's experience also helped to release the inhibition common to mathematicians against saying how things really are, rather than how they would have to be *if* they were to make a neat logical package.

The material which follows, then, is designed to have two complementary functions:

(1) To help the reader face up to what the situation is really like, to what statistics can and cannot do for him, to which burdens of uncertainty and judgment he must shoulder if quantitative procedures are to serve him well.

(2) To point out to him how he can set about getting more out of his data by treating it differently, sometimes by making very simple changes in his practices, sometimes by adapting or adopting some of the newer techniques of analysis.

For the last few decades mathematical and theoretical statistics has concentrated on the search for certainty in the face of uncertainty. (See G2 below.) Rather more of the present account is devoted to the most classical results of this search, techniques for assessing significance and asserting confidence, than I should like. But many pages, I boast, are devoted to better ways to dissect data so as to see what is going on, to techniques of incision, rather than to those of conclusion or of decision. Dissection is the heart of data analysis, and each man who studies data needs to continue to learn new ways of dissection, to master new incisive techniques.

The original draft of this account was written as background for a Seminar at the Center. It was revised and extended after I had returned to Princeton Unversity and Bell Telephone Laboratories. The work for Princeton was supported by the Office of Ordinance Research, U.S. Army under Contract DA 36-034-ORD-2297. Without the support of all these organizations and the atmosphere of the Center, it would never have been written. Those readers who find it helpful should give thanks accordingly.

TABLE OF CONTENTS

INTRODUCTION

The structure of this account is unconventional. And why not? It opens with a list of "badmandments," of unwise statements which most of us can imagine *someone else* teaching to his students, either by word or deed. It may be that some of these are so obviously "bad" that they do not need to be set out and stoned. I hope this may be true, but I fear it is false. (They are not living persons, nor are they living truths, therefore all of us should cast stones, especially those who have themselves sinned.)

From these badmandments it is easy to formulate questions, and to go on and discuss various topics. Readers are encouraged to wander and browse among the discussions. While there is some system in their arrangement, and a small amount of mutual dependence, system and dependence are only there for those, probably a minority, who want to

read through systematically. An epilogue attempts to draw together the branching strands of the discussion. And several appendices contain supporting material.

The original version promised much philosophy, with only a seasoning of techniques. The snowball has grown since then; techniques have been like the wet snow, philosophy like the dry. A fair number of simple or novel techniques are now included, but philosophy still manages to dominate, perhaps barely. Psychology and economics, once almost untouched, have now contributed examples.

PROLOGUE

Once upon a time, a collection of cautionary admonitions of the sort that follow would have been introduced by the words "once upon a time" · · · and a story of the finding of an ancient manuscript in a hidden and cabalistic cache. Today one begins differently, by thanking friends and colleagues for their help and contributions. The admonitions which follow are, indeed evil; they are not mere straw men or scapegoats. They are truly badmandments. For:

(1) I can imagine each of them being supported, explicitly or implicitly, by more people than I would wish.

(2) Following their directions can, and usually will, lead to bad statistics and bad science.

Each reader should ask, as reads a badmandment: "How prevalent is this one in my own thoughts? In those of my immediate colleagues? In my special field as a whole?"

THE GREAT BADMANDMENT

The great badmandment can be stated in all languages and to apply to any situation. In general allegorical language it reads:

IF IT'S MESSY, SWEEP IT UNDER THE RUG.

It is not difficult to restate it in form specially pertinent for behavioral science, and to trace some of its consequences by first stating five

badmandments that flow from it, and then expanding each of these five prime badmandments further. (As in Tables 1 to 5.) Shock effect has therapeutic values, so all these restatements and consequences are rightly put in pungent language. The great badmandment is simple, and of universal application. The first of its prime consequences is rather specifically related to behavioral science, but the others apply widely, providing stings for many investigators in a wide variety of other fields.

THE GREAT BADMANDMENT RESTATED, AND FIVE PRIME CONSEQUENCES

ONLY THREE ACTIONS IN SCIENCE ARE SAFE: TO BE GUIDED BY THEORY, any theory; TO BE SIMPLE, very simple, and TO DO NOTHING, absolutely nothing. (· · · Certainly, you must be safe at any cost! · · · You might miss something? Don't worry; so long as you stick as close as a flea to some combination of the three safe actions, no one will ever know what you missed!)

1. THERE IS NO ANALYSIS LIKE UNTO CROSS-TABULATION. (And the counting sorter is its prophet! · · · It's simple.) (See Table 1.)
2. BE EXACTLY WRONG, RATHER THAN APPROXIMATELY RIGHT. (At all costs, be *exact*! · · · Theory is exact and 'exact' is theory.) (See Table 2.)
3. THE ONE AND ONLY PROPER USE OF STATISTICS IS FOR SANCTIFICATION. (We used statistics, our work is above criticism! · · · Statistics is theoretical.) (See Table 3.)
4. BEWARE EMPIRICISM, IT ISN'T SCIENTIFIC. (And we must be scientific, even if we learn nothing · · · Empiricism can be dangerous.) (See Table 4.)
5. AT ALL COSTS BE RIGID AND SERIOUS; FOLLOW THE STRAIGHT AND NARROW WAY TO ITS INEVITABLE END. (A *scientist* always knows where he's going! · · · He might get in trouble otherwise.) (See Table 5.)

Table 1

The First Badmandment Expanded

1. THERE IS NO ANALYSIS LIKE UNTO CROSS TABULATION. (And the counting sorter is its prophet!)

    11. CROSS-TABULATE till the numbers are *almost* too small, then STOP.
    (No one can criticize you · · · you've done *all* that is *possible*!)

    12. DON'T try to make difficult distinctions; use only two or three cells for each SCALE.
    (If you make too many cells, you can't cross-tabulate enough ways at once. If you try to make distinctions which are too difficult you will put some cases in the *wrong* cells; just think what *that* would do to your *cross*-tabulations!)

    13. NEVER describe a split except by counts or a PERCENTAGE.
    (The human mind cannot understand any other description of a breakdown into two classes!)

    14. CELLS in a cross-tabulation involving less than 10 cases are USELESS.
    (If you can't or won't combine rows or columns to get rid of them, give no information at all for such cells · · · Make the table less useful? Sure, but it will keep some people from getting wrong ideas!)

Table 2

The Second Badmandment Expanded

21. BE EXACTLY WRONG, RATHER THAN APPROXIMATELY RIGHT. (At all costs, be *exact*!)

    21. ALWAYS use "invariantive" statistics appropriate to the scale type* of your MEASUREMENTS.
    (Then no one can question your judgment · · · since you didn't use any!)

    22. NEVER make inferences to hypothetical POPULATIONS.
    (Stop with a real population! Avoid uncertain populations like the plague. Then probability sampling can make your formal inference so tight that hardly anyone will think about the remaining uncertainties, especially the ever-present informal inference from where you really stopped to where you want to go!)

    23. NEVER make any analysis which was not planned before seeing the DATA.
    (You might learn something new and unexpected, but you couldn't put a precise significance level on it, now *could* you? · · · What, preanalyze a small random subsample as a guide to analyzing the whole? But that would be so *unusual*!)

    24. QUANTITATIVE empirical regularities are useless in the present state of our SCIENCE.
    (Further study will always show that they were not *precise*, were not expressed in precisely the *right* form! · · · You *know* it will!)

    25. IF order is the only guaranteed property of your scale, DICHOTOMIZE!
    (Then you won't have used an incorrect scale! · · · Thrown away data? Why I suppose you will, but *that* isn't so important! · · · Made analysis more complex? Not if you only cross-tabulate!)

---

* If it isn't *exactly* an interval scale, it's only ordinal; if it isn't *exactly* a ratio scale, it's only interval. (Cp. D6 below).

Table 3

The Third Badmandment Expanded

3. THE ONE AND ONLY PROPER USE OF STATISTICS IS FOR SANCTIFICATION.

   31. IF a statistical significance test can't demonstrate *causal* relations · · · throw it AWAY.
   (It must be useless, we want *only irreproachable general results*!)

   32. NEVER use statistical techniques to help you find interesting INDICATIONS.
   (· · · Yes, it might be very helpful. But what a *perversion*, the indications might *not* be significant!)

   33. DISTINGUISH, unfailingly and forever, even barely *statistically significant* results from ones that do *not* reach SIGNIFICANCE.
   (· · · You say the underlying strength of relation *might* be the same? But only *one* was *significant*!).

   34. ONCE a number of results are statistically significantly not all the same, believe all apparent relationships among them IMPLICITLY.
   (Mind you, his differences yield an *F* statistically *significant* at the 1% *level*! *Surely* that unexpected difference between A and B must be *real*!)

   35. IF one overall test shows lack of significance, STOP.
   (· · · F-test among means? Sure! · · · Chi-square for contingency table association? Sure, Mike!)

   36. DON'T think, use STATISTICS.
   (Why of course! What else are statistical techniques for?)

   37. ALWAYS use the 5% level of SIGNIFICANCE
   (*Everybody* who is anybody always *does*! Think about what is really appropriate to your situation? How *odd*!)

   38. IF a result is not significant, don't dare show IT.
   (Some poor *fool* might be misled into *believing* it! · · · It *might* be *right*? So what!)

Table 4

The Fourth Badmandment Expanded

4. BEWARE EMPIRICISM, IT ISN'T SCIENTIFIC. (And we must be scientific, even if we learn nothing).

   41. WORTHWHILE regularities always come equipped with a theoretical EXPLANATION.
   (· · · You found a *more* empirical regularity? Forget it. · · · It would have important consequences? Not if it's only empirical.)

   42. QUANTITATIVE measures are most dangerous when they *seem* to behave unexplainably WELL.
   (His fit was too good! No *theory* could ever account for it! It *must* have been an *artifact*!)

   43. DOING anything at all, except nothing (or, perhaps, except exactly that which is conventional) is being ARBITRARY.
   (· · · It might be better? But it would be arbitrary!)

   44. ONCE your description fits roughly, STOP.
   (Never, never study the deviations of observed from described · · · Why, you might find *systematic deviations*! And then *where* would you *be*?)

Table 5

The Fifth Badmandment Expanded

5. AT ALL COSTS BE RIGID AND SERIOUS; FOLLOW THE STRAIGHT AND NARROW WAY TO ITS INEVITABLE END. (A scientist always knows where he's going!)

    51. DON'T try out your proposed data-gathering instrument (questionnaire, record-searching technique, etc.) on a preliminary SAMPLE.
    (· · · Pretest your questionnaire? But you might have to change it!)

    52. DON'T try out your proposed method of analysis on a preliminary SAMPLE.
    (· · · Pretest your analysis? But you might have to change it!)

    53. DON'T admit, even to yourself, that you had to begin with EXPLORATION.
    (It's very improper to work without definite hypotheses · · · Half-hypotheses? What a weird idea!)

    54. A good piece of work is DEFINITIVE.
    (· · · He admits this is only the *first* phase? His work must be *utterly* useless!)

    55. ANY empirical observation must either be considered useless, or taken very, very SERIOUSLY.
    (There is no room for a *middle* ground. It *must* be either *right* or *wrong*!)

    56. DON'T try to find a simple way to answer a complicated QUESTION.
    (· · · He showed one *photograph*? But where were his data? Yes, I know the photograph was *conclusive*, but was it *data*? · · ·)

SUPPLEMENTARY BADMANDMENTS

Table 6 contains some supplementary badmandments, mainly suggested by colleagues. (I owe certain of the expansion of the first five badmandments to colleagues also. Many thanks!) In most cases it is not difficult to trace each of these back to the prime badmandment. We leave this as an instructive task for the interested reader.

Table 6

6. SOME SUPPLEMENTARY BADMANDMENTS, MAINLY SUGGESTED BY COLLEAGUES.

91. NEVER plan any analysis before seeing the DATA. (Why, *who* can tell what you may learn from even *three* cases?)

92. DON'T consult with a statistician until after collecting your data; you would only get confused and DISCOURAGED! (· · · Maybe he could help you get *more useful* data? But would it be *right*? · · · Smith saved *three* years work? But was that *science*?)

93. IT is far, far better to have a large, obvious, but statistically not significant difference, than one that is small, reliable and statistically SIGNIFICANT.
(His differences were *well* established? But look how much bigger Smith's were!)

94. LARGE enough samples always tell the TRUTH. (· · · There wasn't *anything* random about his sample? But look *how many* cases! · · · The Literary Digest Poll failed *dismally*? But that was *so* many years ago!)

95. NEVER tell your statistical consultant about the two most important recent papers in the field of your own RESEARCH (· · · It might help him *advise* you? But he is only supposed to help with *statistics*! Jones's statistician can think about Jones's *problem*? How *odd*!)

96. NEVER try to find out if your population is meaningfully divided into two or more SUBPOPULATIONS. (· · · His data made *much* more *sense* when it was separated on *that* variable? It couldn't be! That variable can't be *important*!)

97. ANY one regression will tell you what you want to know, don't even think of looking at MORE. (· · · He tried various *alternative* regressions? How *odd*! · · · She looked at regressions within subgroups? But why, oh why?)

98. IF $r_{xy.z}$ is significantly different from zero, z can't possibly explain the relation between x and y. (· · · Yes, I've heard of *orthogonal polynomials*! Buy they're just for *curve fitting*!)

99. IF a "before score" goes with each of your "after scores," always analyze the DIFFERENCES. (· · · Use covariance? Don't be foolish! · · · Your before score is too variable? Nonsense!)

100. The significance level tells you the probability your result is WRONG. (· · · Yes, I know the *books* say something else, but I *must know* the probability that I'm *wrong*. · · · Robinson, Sr., *never* tests any null hypothesis except one that everyone knows *couldn't possibly hold*? But his results are *never* significant at the *zero* % level! · · · Robinson, Jr., *never* tests any null hypothesis except one that *must hold*? But once in 20 times his results are *significant*, significant at the 5% *level*!)

QUESTIONS

It is easy to derive corresponding questions, stated in sober language, from most of these badmandments, and to supplement these questions with either answers or references to the discussion. (It is convenient to give each question the number of the badmandment to which it corresponds.)

11. How can we go further than by cross-tabulation? (See E and F.)

12A. Will we then need to restrict ourselves to few categories along each scale? (See D and F.)

12B. What do errors of classification really cost us? (See D1.)

13. What other ways of describing splits are useful, and why? (See D2 and D3.)

14. What use can be made of data from less than 10 observations per cell? (See T4 and F2.)

21. What are the pros and cons about the use of "sophisticated" statistics like means and standard deviations when the "measurements" are on a scale where only order is definite? (See D5.)

22. What are some of the pros and cons about the point where formal statistical inferences should stop? (See B.)

23. Is it wise to let a body of data guide its own analysis? (See C3.)

24A. Are quantitative empirical regularities valuable? (See C1.)

24B. Are quantitative empirical regularities shorter lived than theories? Or longer lived? (See C1.)

25. Why is it unwise to dichotomize data available on a more extended scale? (See D.)

31. How may causal relations be established? (See A1 and W.)

32. Can statistical techniques be used to glean interesting indications from data? (The answer is "certainly!")

33A. Why do users tend to erect such a rigid wall between results which are statistically significant and those which are not? (See C2.)

33B. What are some of the fallacies encouraged by such walls? (See G1.)

34. Can we usefully come to more diverse and useful conclusions about the mutual relations of several quantitative results than "they could be alike", and "they are significantly different, believe in all appearances"? (See G4.)

35. Can we honestly do more than make one overall test of significance? (See G2 and G3.)

36A. Is thinking proper? (The answer is "yes".)

36B. Can we learn to think more clearly? (The psychologists should answer this!)

37. Why are there tables for more than one level of significance? (section not yet written)

38. How can there be any interest in results that aren't significant? (See G1.)

41. Are purely empirical regularities worthwhile? (See C1.)

42. What good can come of unusually well-behaved quantitative measures? (See C1.)

43. What really constitutes being arbitrary? (See A2.)

44. Is anything to be learned from the study of residuals? (See F4 to F6.)

51. Does it pay to pretest questionnaires and other data gathering instruments? (The answer is "yes".)

52. Does it pay to pretest methods of analysis? (See F6.)

53. Is exploratory inquiry efficient science? (The answer is "yes".)

55. Can empirical observations be usefully taken as working tools for later sharpening? (The answer is "yes".)

56. Does it pay to find a simple way to answer a complicated question? (The answer is "every time".)

A. GENERAL ATTITUDES

A1. HOW MAY CAUSAL RELATIONS BE ESTABLISHED?

The answer to such a question cannot be derived from mutually-agreed-upon hypotheses by formal procedures. The most that can be sought is a point of view buttressed by a more or less convincing analysis, and by illuminating examples. This we shall try to provide.

The point of view is simple. The establishment of a causal relation always requires two elements, one empirical, the other theoretical. The *empirical* observed regularity or experimental result has to be such that its occurrence is *theoretically* impossible unless "A caused B". Both elements, the empirical and the theoretical, are essential. Neither alone can *establish* causation; both are required. An empirical result alone can *suggest* causation, and this suggestion can be strengthened by theoretical considerations which make it less and less likely that the particular empirical result would ever occur unless "A caused B". (These theoretical considerations must thus tend to rule out such possibilities as (i) "B caused A" or (ii) "something else caused both A and B, or caused B and was associated with A".)

If this view be sound, it has very important implications about "The Statistician's Burden". For, if it be sound, statistics has no responsibility beyond what we might call empirical projection · · · the inference from certain empirical observations to what would happen, empirically, *if* observations or experiments were made on a much larger scale. Such a situation may be contrafactual, but is (or would be) empirical. Consequently, it would, by itself, be without causal content. The inference from such a "projected" empirical result to causation is

then a responsibility of theory, and its purveyor, the subject-matter specialist. (As a theoretical concept, "causation" seems to me to be unequivocally useful, even when all its misuses are allowed for.)

This avoidance of a heavier "Statistician's Burden" may not seem important, especially in the behavioral sciences. Yet when Hanan Selvin's recent paper in the *American Sociological Review* attacking the use of significance tests in sociology (Selven 1957) is examined, the basic motivations for his evil impressions appear to be two:

(1) Some sociologists (along with some statisticians and some members of all statistics-using professions) misuse significance tests, and

(2) significance tests cannot establish causation.

Of the two, the latter appears to bother him the most.

Why would one choose to adopt this point of view? It seems to me that all the really clinching arguments as to causation in particular situations come down to saying "there was no way for B to affect A, hence (barring some C as the cause of both) the observed accompaniment (whether uniform and constant or merely statistically excessive) of A and B must be due to causation of B by A." In such processes two general principles are applied in many instances, namely:

I. It is impossible for an event occurring at a later time to cause an event occurring at an earlier time.

II. It is impossible for the (possibly concealed) factors; which may determine, to a lesser or greater degree, outcomes which will occur for specific experimental units to affect the selection of units for specific treatments (including the "control" treatment) when this selection is made by rolling dice, shuffling cards, reading out random numbers — or even, as some appear to feel, quite without empirical justification, by the "random" judgment of an experimenter.

Most statements of established causation in the physical and biological sciences involve one of these principles. Thus no one doubts that the change in a star which accompanies its great brightening as a nova causes the shell of luminous gas *later* observed to surround the star. And no one doubted that mosquitoes carried yellow fever once those selected to be, and actually, bitten by infected mosquitos contracted the disease, while the remaining subjects did not.

A somewhat less general statement of a similar position, in the area of survey technique, has been made by Hyman (1955), who says: "The notion of explanation provides an analytic basis for defining clearly a causal relationship between two variables. *If the partial relationships never disappear, even when every conceivable antecedent test factor is introduced, then the original relationship is a causal one.*" (Italics Hyman's; see Section D3 for a cautionary example, however.)

It is very illuminating, in passing, whatever one's views about parapsychological powers may be, to consider what would be the effect on one's judgment of causation if we admitted the reality of such powers. If an experimenter had precognitive clairvoyance, for example, and could know just which subjects were going to contract yellow fever, he could arrange for these subjects to be bitten by "infected" mosquitoes. The experiment would then offer no evidence of causation. (Designers of experiments may find the problem of distinguishing "immediate clairvoyance" from "precognitive telepathy", given that one and only one exists, quite interesting. It is rumored that one solution is offered in Carington 1945.)

Similarly, if R. P. Feynman's model of positons (positive electrons) as ordinary electrons moving backwards in time (in his model, pair production or annihilation are just U-turns) should grow into a physics in which influences could travel backward in time, how would we know that the later gas shell did not cause the earlier events in the nova?

In this brief discussion, we have not tried to say all that we might about either significance (about which somewhat more will be said in G below) or the establishment of causation (to some aspects of which we return at the end of W). And we have not really touched on the definition of causation (see Wold 1966 for one view) or on why it is useful (Tukey 1954 may shed some light on this). But the basic idea underlying Selvin's criticism seemed so important, and so little discussed, as to deserve special and early notice.

A2. WHAT CONSTITUTES ARBITRARINESS?

What procedures of acquiring data, of processing data, of interpreting processed data, are arbitrary? This question is at least as broad as the last one. And no neatly-packaged answer is easily available. About all we can hope to do is to exhibit and discuss some very poor choices of what is bad because it is "arbitrary".

Sometimes "arbitrary" means merely "not the way we are accustomed to do it." Thus an engineer used to volts, amperes and ohms is quite likely to regard the description of the strength of

electrical signals in "decibels above reference" as quite arbitrary the first time or two he meets this usage (though he soon learns its virtues). Similarly, the casual counter of the numbers of fleas or of mites on rats is likely to regard the use of such forms of expression as

$$\sqrt{\text{number of fleas}}$$

or

$$\log (1 + \text{number of mites})$$

as very arbitrary, though he too is likely to come to see some of their advantages. Clearly, however, unfamiliarity can be described in other words, and does not need "arbitrary" as a further synonym.

Sometimes "arbitrary" means "not in one of the accepted patterns." Thus if a single rat has run a maze 30 times it is usually not regarded as "arbitrary" (though it is most usually inappropriate) to assume either that one has 30 independent observations, or that one has one observation (one rat, one observation). On the other hand, it *is* often regarded as "arbitrary" to assume that rat-to-rat variations are one-half the size of trial-to-trial variations for an individual rat. (Such an assumption would make 30 observations on one rat the equivalent of about 4.4 individual trials on separate rats.) It is not "arbitrary" to assume that rat-to-rat fluctuations are very, very small — or very, very large — compared to trial-to-trial fluctuations, but it is "arbitrary" to assume them to be one-half as large. Just how, and why, is $\frac{1}{2}$ more "arbitrary" than 0 or ∞? (To say that it is not an accepted pattern seems not to be enough.)

Mainly, I believe, because it is humanly possible to forget, actually or formally, one source of variation whenever 0 or ∞ is assumed, which involves acting as if one source of variation were negligible as compared to the other. And the act of neglecting something is so close to doing nothing as to be *thought* "not arbitrary". On the other hand, when $\frac{1}{2}$ is assumed, both sources of variation must be recognized, and consideration of the possibility that the ratio might be 0.3 or 0.7 (instead of 0.5) recurs, whether one likes it or not. While this suggestion makes this attitude psychologically (or perhaps psychiatrically) more understandable, it does not make it a bit more logical, nor does it make it a bit more effective as an aid to gaining knowledge.

Thus, as this instance suggests, the "not-in-an-accepted-pattern" sort of arbitrariness may have been generated by what to me seems to be the greatest fallacy of them all: ***the belief that doing nothing cannot be arbitrary.*** Just how this view comes about is not really clear. To what

extent does it derive from the view that "exactness" is essential at any price? To what extent are these two views only seemingly convergent? To what extent do they derive from some underlying, unperceived line or lines of thought and feeling? Or do they, perhaps, come from a misconception of the meaning of the word "exact" in that laudatory, sought-after, brightly shining phrase "The Exact Sciences"? (To describe "exact scientists" as a reference group for some behavioral scientists would be to understate the strength of their feelings, but to take off the quotation marks would be to make this statement wholly false.)

A chemist is an "exact scientist", particularly if he is an analytical chemist. He weighs, he measures, he determines. And how does he do these things? He weighs on a chemical balance, using a set of weights. Does he do nothing about the weights, thus avoiding arbitrariness? Not at all. His first task is to calibrate his set of weights, determining corrections to their nominal values so that he may thereafter weigh more precisely. Are these calibration corrections themselves to be regarded as "correct"? Surely not. He recognizes that redoing the whole calibration would lead to slightly different corrections, but he has reason to believe that his are good enough to be useful. In other words, the corrections to his weights are "arbitrary" in the sense that they are not supposed to be ultimately exact. They do, however, belong to a club of lower prestige but higher usefulness, for they may be wisely thought to be "either good enough, or about as well as we can do."

The chemist also measures liquids, and titrates to various end points, some defined in terms of "neutrality". He measures liquids with a burette or a pipette. And his first task is to calibrate these devices. He titrates to neutrality with an indicator. And he arbitrarily chooses that indicator (or that color of a universal indicator) which experience shows gives the best results. (Then he standardizes his titrating solution on a known sample.) Throughout he proceeds by making corrections and adjustments to get the most precise and useful value. None of these adjustments are "exact", all are "arbitrary" in the sense that, if done over, they would be different. But ask any chemist if it would not be better to omit them, to be "exact" by not being "arbitrary".

The nature of "The Exact Sciences" is that they are full of "corrections", "art" and what might even appear to be "folk-wisdom", especially when one is concerned with the practice of measurement. Other fields cannot hope to become "Exact" with a *capital* E by abjuring good quantitative judgment, or by abjuring empirically sound adjustments, or by abjuring "arbitrary" corrections. (Such actions can only lead to "exactness" with a vanishingly small "e", and, inevitably, to a vanishingly small effectiveness.)

B. THE DIMENSIONS OF INFERENCE

B1. THE TWO SORTS OF GAP

To ask a behavioral scientist "What are the dimensions of inference?" would be to offer opportunity for many diverse replies, since "dimension" comes closer than any other mathematical term to being all things to all people. In the title of this section, however, the reference is to the dimensions of a piece of lumber: to length, breadth and thickness. All three are important to the builder of wooden structures. All are important to the user of inferences (meaning formal or informal ways of passing from the particular toward the general). But, you may properly say, how can an inference have a length, a breadth, or a thickness; admittedly these words must be used by analogy, but by what analogies?

In his book on *The Design of Experiments* (Fisher 1935ff), a book whose understanding requires some statistical background and whose reading repays frequent repetition, R. A. (now Sir Ronald) Fisher points out the advantages of broader bases to inferences. As exemplified by the advantages of detecting a phenomenon in 5 widely different cultures, rather than in 5 West African tribes, this idea is familiar to behavioral scientists. As exemplified in cross-tabulation for the purpose of showing that the effect still occurs in each stratum, it is likewise familiar. As exemplified in analyses where breakdowns are carried so far as to require some sort of recombination before interpretation (see F2 below), and in other analytically sophisticated expressions, the idea is not so familiar, is not nearly familiar enough.

It is not unnatural to describe the extent of the data involved as the "thickness" of the inference. If we think of the inference as a bridge which helps us on our way, then both breadth and thickness help to provide strength. (Indeed, an excess of one cannot make up for a deficiency of the other.)

But there is a further dimension of more or less formal inference, one all too often unrecognized — its "length". It is too easy to forget that scientific (or practical) inferences have to span gaps far wider than any statistical bridge (or any possible formally logical bridge) can reach. In fact, it would seem that it usually does not even have to be forgotten, having never been consciously recognized. Thus we shall mention very diverse examples, in the hope of synergistic arousal.

In the practical applications of the exact sciences we see this gap recognized daily, though we do not think of it as such. The chemical engineer's classical chain of development — laboratory, pilot plant, semi-works, full scale — was (and generally remains) an admission that no amount of small-scale testing in the laboratory would settle what would happen in the works. (Thickness could be provided by many tests, many sorts of breadth could be provided by changes in reaction vessels and manipulations, but length remains insufficient to bridge the gap.) No one would have expected the atomic bomb to go from Los Alamos to Hiroshima without a stop at Alamogordo. (In fact, the surprising thing was that *one* large-scale trial was enough.)

We are "future-oriented" in all fields of science and technology, we study "the present" (the recent past) and "the past" (the distant past) with the hope of foreseeing, and perhaps even guiding, the future. ***Purely statistical considerations alone can never suffice*** for inferences from the past (distant or recent) to the future (at least not until time machines are available). For we cannot draw samples from the future. We can make statistical inferences from what was observed in some sample of the past to larger aspects of the past. We may even, indeed, make inferences to "might-have-been" pasts (the latter is the most important function of much of modern statistics). But we may go no further by purely statistical arguments. Only theory (itself held on faith) can guarantee that the future will resemble the past. (The "laws of nature" *may* be due for a sudden change at 4:23 a.m. on the next Saturday 29th March "that ever is".) This gap between past and future is common to us all, both personally and by disciplines. Its recognition is of little importance in itself, for there is little that we can do but to recognize its presence and then press on, trusting our faith in the continuities of nature. It is mentioned here, however, to help throw light on less extreme gaps of a somewhat similar nature, gaps enforced not by "Time's Arrow", but by the extent of the data actually available to us, which are the subject of the next few examples.

B2. A BIOLOGICAL EXAMPLE

The biologist studying genetics in insects has habitually used the little fruit-fly *Drosophila*. And in this genus he has used certain species, often working with a single laboratory colony of a single geographic race of a single species. In doing this he has probably not been unwise, but he has introduced many gaps of a sort not to be crossed with statistics. No matter how thick the inference, no matter how many flies are raised and classed, just so long as all the flies come from a single

geographical race of a single species, the *purely statistical* inference clearly cannot extend outside the family of which *Drosophila* is a genus, nor outside that genus, nor outside the species studies, nor outside the geographical race from which the colony sprang (and often not outside the colony). Yet the biologist's interests are not bounded by these limits. Few biologists would be modest enough to feel that they were studying fly genetics. Most would feel that they were studying the fundamental mechanisms of genetics. Of the span of inference from the specific laboratory fruit flies to all flies, to all insects, or to all life, only a short span can be statistical, most has to be biological.

But this has caused little confusion about the contribution of statistics to the inference. A man claiming a new genetic phenomenon in *Drosophila* dare not say merely that more than the previously expected 50% of flies show this characteristic on the sole basis that more than half of the flies he examined showed this characteristic. (After all, there is the "66.7% cured" of the medical article traditional among statisticians, namely 2 out of 3.) The claimer will be forced to consider his flies a sample, even if he has studied every fly in his own colony, even if any larger "population" from which these flies might be regarded as a "sample" were purely conceptual and never existed as such. Here the policy is well established.

When we deal with people instead of flies, the situation is not so clear (perhaps because the investigators do not live many times as long as the subjects). The next few pages will summarize a human illustration, homely in one sense (though we trust not in another).

B3. A BEAUTY CONTEST

We now wish to discuss the same problem in terms of a bathing beauty contest at a seaside resort. Let us suppose that some 25 girls are judged by a panel of 300 men, drawn at random from the adult male population of the resort. Suppose further that each judge rates each contestant on a scale from 0 to 100, and that we are concerned with the average rating which would have been given to a contestant by all male residents. For each contestant, the determined (i.e., that which is to be pointed toward by an appropriate summarization of the data) is this average, a typical value of the population of scores which would have been given by all adult male residents. The natural choice of the corresponding determination or estimate is the mean of the 300 scores actually given that contestant.

Comparisons between individual girls are clearly of interest. (As usual, simple comparisons are important.) But, given the data, it is reasonably certain that someone (probably several people) will wish to make comparisons between redheads, blondes and brunettes. He, she, or they will almost inevitably calculate the mean score of all redheads, the mean score of all blondes, and the mean score of all brunettes, and start to intercompare these mean scores. What difficulties must now be faced that are likely to be overlooked?

He (or she) will undoubtedly be concerned with some generally expressed question, such as "Do men prefer blondes?" It is most unlikely that his (or her) curiosity extends only to the particular girls who participated in the particular contest (particularly if hair dyeing or tinting may be in question). Now there is little doubt but that the blondes who entered this contest are *not* a random sample of blondes, that the brunettes are *not* a random sample of brunettes, etc. (It might be possible to get a random sample of adult males to act as judges, but hardly conceivable that a random sample of girls would be willing to become contestants.) In fact, we can plausibly say more. Is it not a fact that the relative number of redheads in such a contest is greater than in the female population at large (of appropriate ages)? If so, then either the selective forces of recruiting contestants must operate differently for redheads, or the selective forces that determine hair color must operate differently for potential contestants. In either case, some systematic effects of a difference, or of hair differences, are to be anticipated.

But our protagonist will overcome this difficulty, probably in the only reasonable way, namely by deciding that he wishes to compare the average adjudged beauty, not of all girls of a given hair color but of those who "might have been contestants in a similar contest." (Note carefully, not merely those who *actually* entered *this* contest.) To the sampler of well-defined populations by modern methods of probability sampling, this may seem an atrocious action. He might say: "The 'populations' now being considered are not definite enough! There is no trace of a list or frame covering all individuals. You cannot even tell whether a particular girl belongs to this 'population' or not. Probability sampling was not used to select the sample, so the use of formal machinery based on random sampling is entirely improper."

To be sure, there is a real uncertainty here, but it is not novel, not unusual, and, in the writer's judgment not too serious. Let him try to explain why.

B4. EXAMPLES FROM PHYSICAL SCIENCE

This problem of the uncertainty of any possible population reference has been faced elsewhere, and experience has often shown that it is better to make inferences to such an *uncertain* population than to tie oneself down to a particular sample. Modern statistics developed in closer relation to agricultural experimentation than to any other single field of science or technology. Agricultural experimentation is affected by weather — most seriously affected. And is the weather of last year, this year and next year a sample from a *well-defined population* of annual weather patterns? Is it a *random* sample drawn with known probabilities? Only the briefest analysis of historical data is required to show that the answer is irretrievably "no" to both questions. (Even less time is needed to discard that course of perfection which says: "So, you're interested in the average behavior of these crops during the next 25 years since you need to make recommendations to farmers which will be valid over that period. Why not select 5 of these years at random, say one from each of 5 five-year strata, and plan to run your experiments in these selected 5 years?") What practice is in regular use in the assessment of agricultural field trials? What has experience taught? A very simple rule of thumb: "Treat the years you have as a random sample of years from a population of 'similar' years; this is the best you can do!" Here there has been extensive experience; moreover when the years at hand were treated in some other way, experience has often been bitter.

When statistical techniques are applied to experiments in chemistry and chemical engineering (and of all the technologies today, chemical industry makes the greatest, and most rapidly growing, use of statistics) the "samples" often arise by doing the same thing twice, three times or more. A chemist may analyze three aliquots of a liquid sample. Three different chemists may each analyze a sample from the same batch (they may or may not be in three different laboratories). Three replicate fermentations may be run in the laboratory 50-gallon fermenter. Three experimental runs may be made, one in each of the plant's three 20,000 gallon fermenters. Three experimental runs of four days each may be made in the refinery's one big catalytic cracking still. And so on. In each case the three observations will be treated as a "sample". And in each case the "population" will be impalpable, unlistable and unframable, and *uncertain*. In every instance, however, the inference will have been to a "population of similar runs." The uncertainty of the precise nature and identity of such populations has not inhibited or devalued the developing use of statistical techniques in the chemical industry, where experimental conclusions lead to plant-scale trials, and

where, contrary to experience in some academic fields, mistakes almost inevitably come home to roost. "Populations of similar runs" has proved to be a valuable concept.

Consider now a third example. If you were an astronomer studying the average behavior of certain novae, and found that good spectroscopic plates were available for 17 instances (new instances being expected at a rate of about 1 nova every 2 years), would you treat these 17 "exploding stars" as a sample? As an entire population? And to precisely what population would you refer them?

When subgroup means cannot bear the weight of inference to any well-defined population, it may be that ***logically*** they will not bear the weight of the inference to an uncertain population, but practically it is almost certain to be best to use them as the bases of such inferences to uncertain populations.

B5. A SOCIOLOGICAL EXAMPLE

Consider next a sociological example, mentioned in a Center seminar. Sociologists interested in the process of professionalization studied the classes attending a given medical school during a particular year. Relations between students mainly in the same class, but also between classes, were of considerable importance. Let us suppose for illustration, whether or not it be the case in fact, that certain processes of professionalization were peculiarly distinctive in the first year, freshman class. How are we to regard the available evidence? It relates to all the students in the freshman class, the freshman class of this particular year, in this particular school, a school which trained for the particular profession, medicine! We want to learn about professionalism, not just in medicine, not just in this school, not just in this academic year. Clearly much of this span cannot be covered by statistics. But some of it can be covered. From a general point of view, it is equally as desirable to interpret these medical students who were there as representative of the hypothetical population of medical students who might have been there as it is to interpret those *Drosophila* who were there as representative of the hypothetical population of *Drosophila* who might have been there. (There are some special considerations in this sociological example; they will be discussed shortly.) It is clearly desirable, and often essential, to extend the statistical part of the inference as far as we can. There will still be a wide enough span left to subject-matter faith alone.

If we can agree on the general principle, what of the implementation? What are the special considerations which arise in the two cases? With *Drosophila* it may be important that individual flies are not unrelated; proper analysis of the data may require treating the data in terms of "progenies" or in terms of the groups of progenies raised in separate "bottles." With medical students, the situation is somewhat more complex. While it is probably true that the student-school selection process does not select students independently (and it is undoubtedly true that different schools have different curricula and policies), a much more serious process begins once the students arrive at medical school, once they begin to become a "Freshman class". This process involves much interaction between persons, and the nonrandom development of many relations. (As one example, note that the number of persons classifiable as "*the* most important opinion leader" is not distributed in various Freshman classes as if derived by random sampling from a population containing a certain proportion of such persons. Its average value is presumably greater than 0.5, but since at most *one* person can be the leader, the values 2,3, . . . can never appear.) Thus we have some practical difficulties in making inferences from this "sample" to a hypothetical population, difficulties connected with the determination of the sample size. For some purposes, the sample size may be the number of students, for others we might do reasonably well with the number of clearly recognizable friendship-groups among the students, but for many purposes the sample size is *one*. (One class, one sample!)

B6. SAMPLES OF MORE THAN ONE

From the statistician's point of view, a sample size of one is a serious drawback. In the medical school example, he must ask why the hypothetical study did not follow the professionalization of, say, a random third of the students in each of three medical schools. One sociological answer can be anticipated. It is that "we sociologists study groups as wholes; if we only studied samples we would miss the most important things we are to study." But the force of such an argument is quite limited. If it be feasible to study a sample of individuals, each with his interconnections, such arguments have no weight.

The physical analogy which arises in sampling the out-of-doors plant of a telephone company is interesting, if not too close. The convenient sampling unit is the telephone pole, but poles constitute a minor fraction of the plant. Open wire, cable, cross-arms, guys (and coils, drop loops, and push-braces) are all involved. The solution is

simple. We make up a sampling unit by including with each pole all the equipment carried by or attached to it, thereby including all the interconnections running out from it.

And if it be argued that such sampling of individuals with all interconnections costs more per medical student studied, the answer is immediate. Even if only *one-sixth* of the students in each of *three* freshman classes are studied, there will be a sample size of three for many purposes where before the sample size was one. And while there is much to Milton Friedman's maxim that: "You can never reduce the variance [of the sample mean] as much again as when you increase the sample size from zero to one!", it is equally true that: "You can never reduce the variance so much again as when you increase the sample size from one to three!"

This general point is recognized by sociologists. In a methodological note appended to *The Student Physician* (Merton et al. 1957, p. 304) it is stated that, when a pattern has been found in one medical school: "We consider this a valid result only if the same pattern is observed in a second medical school, or in the same medical school at another time." This statement occurs as part of a discussion of why the authors do not use "significance tests," where those words are taken as meaning conventional tests for counted fractions based upon random sampling of individuals. It would seem that insofar as this aspect of the discussion is concerned, the authors are advocating, not the avoidance of significance tests, but the use of *correct significance tests* based on honest replication (though apparently at what statisticians might consider rather loose significance levels).

B7. ATTITUDES AND CONSCIENCES

We have seen through example something of the need for the use of hypothetical populations, for the shortening of the nonstatistical part of inference by the lengthening of the statistical part. We have seen, especially in connection with sociology, a little of the discomfort and complexity which comes from facing up to the real difficulties of experimental and observational inquiry. We shall later, in G below, have more to say about the problems associated with tests of significance. Here is the place for some broad general comments.

Attitudes to the broad family of questions we have opened up are strongly influenced by views as to the purpose of statistical inference. As we shall try to point out in G2 below, there have been many in every decade of inferential statistics who tried to use statistical techniques as machines for grinding up uncertainty and making

certainty out of the grist. Those who hold such views tend to regard unspecified and unspecifiable populations with disdain and fear. On the other hand, those who, like the writer, look upon statistics as a tool to help us deal with the simpler aspects and kinds of uncertainty, tend to regard extension of statistical inference to an unspecified population as a way to command and control yet one more partial aspect of uncertainty, as serving a very useful function, hence as a definite good.

In the face of the systematic errors which inevitably accompany randomizable errors in every field of science and technology, the writer sees but one view that he himself can take. He cannot insist that others do likewise, since it is clear that everyone ought to make up his own mind about what standard of intellectual honesty, for each individual and for each field, will best support and facilitate progress and sound understanding in the field in question. (And if a man considers instead what standard of intellectual honesty will best support his own professional advancement, we must lay the blame upon his social and professional environment.)

Wherever along this scale a particular user of statistical technique stands, he dare not confuse the ***sampled population,*** which may have to be unspecifiable, with his ***target population,*** itself rather too often unspecified. He must recall that a particular farm, even more a particular field on that farm, even if observed in all kinds of "years", may not be typical (and usually is not typical) of a county, a state, or a country. Similar cautions hold for single ore bodies, single medical schools, single strains of Wistar rats (Williams 1950) and particular classes of students at particular colleges.

Each of us has a right to make short inferences, so long as this is done knowingly, and the remaining gap is recognized. For my own part, I find the concept of a hypothetical population and the making of explicit inferences to such populations not only useful and proper, but important and probably essential to progress. Thus I believe that good practice in a wide variety of fields, including those of behavioral science, involves inferences to hypothetical populations. Some of my statistical colleagues will disagree. The ultimate decision must be made by the scientific consciences of many, many individuals.

C. ATTITUDES TOWARD ANALYTICAL TOOLS

We come now to more specific but not highly restrictive questions of attitude; attitudes toward the empirical and the quantitative in analysis, toward the purposes of formal statistics, and toward allowing the data to guide the course of its own analysis.

C1. THE QUANTITATIVE AND EMPIRICAL IN ANALYSIS

When we think of analyzing data, we usually find the quantitative aspects and the empirical aspects of our analytical techniques entwined in our thoughts. Such aspects are "empirical" when they have arisen more or less directly from contact with data (perhaps from the body of data under analysis, but somewhat more probably from some earlier or more extensive body of more or less similar data) rather than from suggestions by theory or unmitigated "common sense". (Once properly mitigated, common sense is an extremely valuable commodity; but in too raw a state it can be misleading and even dangerous.) Such aspects are quantitative when the resulting comparisons (and analysis of data always involves comparisons with something, if only with alternative anticipations) are expressed quantitatively, expressed not merely in terms like "more than", "equal to" or "less than" but rather in terms like "3 feet higher", "15 points lower I. Q.", "half-way between B and C".

One explanation for this intertwining is so simple as to arouse doubt. It is this: So long as we only concern ourselves ***exclusively*** with "greater" or "less", any reasonable mode of expression "works" as well as any other, and contact with data, even if extensive, has little effect in teaching us how to learn more from similar data. If this be the correct explanation, however, why is there apparent in some areas of the behavioral sciences a miasma of suspicion directed toward the quantitative and the empirical? Somehow there seems to be a feeling that the introduction of such aspects of analysis is dangerous (and I think *not* just threatening) and that results so obtained are piddling and useless. Sometimes these feelings seem to be justified by a reference to the present state of physics, that prototype of exact science, where it is apparently believed that theory and common sense do all the suggesting. But this is a false analogy in many ways. Any meaningful analogy must relate to physics as it was when its state of development was the same as that of the behavioral sciences considered today. And that would be a long time ago, when, in physics, empiricism was rampant and theory minimal. Moreover, even today physics is not unempirical. It is not so that theory precedes careful quantitative measurement. A few examples are easy to give:

(1) The empirical study of "Mach stems", "Mach reflection", and "irregular reflection" as major (quantitative) experimental phenomena of shock wave behavior was very active during World War II. At last reports, theory had not yet caught up with experiment.

(2) The classification of stars into spectral types, and their location on the Russell-Hertzsprung diagram was an important and purely empirical business of observational astronomy for decades. Only today, with a foundation of nuclear physics at hand, are Schwartzschild, Hoyle, and the big electronic computers starting to develop a theory of stellar evolution.

(3) The measurement of the exact wave lengths of spectral lines went on for decades as a purely empirical matter. And when it was learned that using the reciprocal of the wave length made more sense, first because differences between reciprocal wave length were repeated at various places in the same spectrum, and then because the many observed reciprocal wave lengths could be described, empirically, as differences among a smaller set of numbers, empirical work was stimulated but remained empirical. The first theoretical explanation, the Bohr atom, came much later. (And the traces of this history remain today. The shells of electrons are divided, though today this merely means dividing wave functions into families, into those which are "S", those that are "P", those that are "D", etc. — an order with no apparent alphabetic sense. Why? Because the empirical spectroscopist, long before even the empirical discovery of atomic energy levels, had empirically classified spectral lines into families called "sharp", "principal", "diffuse", etc. — a classification made and used three or four decades before there was any corresponding theory.)

As another piece of evidence, I relate a complaint about physicists made by an engineering acquaintance a few years ago: "We have a physicist in the group, but he isn't much help. We told him about a particular situation where the observed results didn't agree with simple theory. First we told him how they deviated from this simple theory. He said: 'Oh yes, it must be that the G is H-ing the K which causes L, etc.' Then we found that we had slipped, and had to go back and tell him that the deviations were in the opposite direction. He spoke up just as fast, saying: 'Well, in that case, the Q must be R-ing the S which causes T, etc.' He can explain anything! How do we get help from him?" Clearly they were dealing with a physical situation where, even today, the empirical precedes the theoretical.

The history of physical science is full of places where *one* precondition of the development of an effective theory was the recognition of an empirical regularity in quantitative terms. Why should matters be otherwise in behavioral science? The abstract of a recent paper entitled "Iterative Experimentation" begins as follows (Box 1957):

"Scientific research is usually an iterative process. The cycle: conjecture-design-experiment-analysis leads to a new cycle of conjecture-design-experiment-analysis and so on. It is helpful to keep this picture of the experimental method in mind when considering statistical problems. Although this cycle is repeated many times during an investigation, the experimental environment in which it is employed and the techniques appropriate for design and analysis tend to change as the investigation proceeds."

"Broadly speaking, one or more of the following four phases can be detected in most investigations:

(a) a screening phase in which an attempt is made to isolate the important variables;

(b) a descriptive phase in which the effects of the variables and the positions of interesting regions of the space of the variables are empirically determined;

(c) a phase leading from (b) to (d);

(d) a theoretical phase in which an attempt is made to understand the actual mechanism of the process studied."

As a consequence of empiricism leading theory in an iterative cycle, successive theoretical explanations often form a nested structure, each new one explaining all that the previous one did and more. Changing the theory need not require changing familiar empirically well-established facts. The evidence of the best-established sciences thus shows that (numbers refer to questions in Table 2 and Table 4):

Quantitative empirical regularities are likely to be most valuable (24A).

Empirical regularities are longer-lived than the theories which their recognition generates (24B).

Most valuable information from observation starts as purely empirical regularities (41).

When some quantitative measure seems to be behaving much better than any theory would suggest (e.g., the reciprocal of wavelength of spectral lines), it is best to push on hard, to use it more widely and more deeply (42).

C2. THE ROLE OF STATISTICS

There are almost as many views of the proper purpose and role of statistics as there are definitions of statistics (and one article collected

531 of these; reference lost to me). Four main purposes for statistical techniques of analysis seem reasonable and important, however:

(1) to aid in summarization;

(2) to aid in "getting at what is going on";

(3) to aid in extracting "information" from the data; and

(4) to aid in communication.

Use for each of these purposes is at least moderately widespread, but all too often an individual use may not be recognized for what it really is.

The uses of statistical techniques in summarization are familiar to most of those who deal with extensive data. A few, who have been exposed to *over*emphasis on modern mathematical statistics, may have allowed summarization to hide so far behind testing, significance and confidence, etc., as to lose sight of it completely. But they can learn easily, either from colleagues who summarize, or from elementary texts. It would be inappropriate to emphasize this class of uses here.

The use of statistical techniques to aid in "getting at what is going on" is another matter. Such simple devices, today to be found profusely sprinkled through books on "general statistics", as typical values (once horribly miscalled "measures of central tendency"!), measures of spread, and measures of association, were once fresh new tools for cutting into and pulling apart quantitative messes. Today the discussion of means, medians, and their relatives; of standard deviations, mean deviations, interquartile ranges, and the like; of coefficients of correlation (product moment, Spearman, Sheppard, tetrachoric, or Kendall) and association; all this seems "old hat". For this there appear to be two reasons.

First, the essential ideas of using typical values, measures of spread, and either measures of nonindependence or of correlation proper as elementary tools of entry into quantitative messes, as incisive techniques, have become part of the tool kit of almost every worker. These ideas may be used with more or less skill, with more or less fluency, with more or less success, but they *are* used. They have lost the exciting aspects of novelty.

Second, new tools for cutting into quantitative messes have been developed. By and large their incisive features have been carefully disguised, and their discussion has been separated from that of the classified tools. The new tools have been disguised by association with experiment, by association with the formal procedures of statistical significance and confidence, by association with careful discussion of what procedures are "best", by association with heavy mathematics. No one of these disguises is necessary, though much thought, ink, and

paper may have to be used to expose these concealed weapons and make them widely useful. Though most users of the old tools are unaware of that fact, there are newer and sharper tools with which one can often cut more deeply and more neatly.

If "descriptive statistics" had been called "incisive statistics", we might have avoided some of this separation of the new from the old. After all, "mere description" does not sound very respectable.

While we shall return to some of the newer tools below (in D, E, F, G and H), we must here urge the practitioner to always examine a new statistical technique, even a highly mathematically packaged one, and ask: "What new sorts of incisions into quantitative messes can it make? What part of it is essential for incision alone?"

The third purpose, extracting information, has been well advertised by the mathematical statisticians. This is natural, and from their point of view fitting, since it is in this connection that mathematical problems which are both interesting and soluble arise most easily. The idea of "squeezing the data" is not unpleasant to the investigator who worked hard to get it, and his cheers have tended to urge the mathematician onward. While there are, as always, dangers of overselling results based on overnarrow hypotheses (such as efficiencies correct for exactly normal distributions and very misleading for *nearly* normal distributions (Tukey 1960)) and of slowness in breaking out into important new areas (such as how to ask of the data in what sort of framework it should be analyzed), the work toward this purpose has on balance been useful and well received, as well as being well advertized and widely recognized as a proper aspect of statistics.

The fourth broad purpose is another which has had little recognition. This is unfortunate, since it is intimately connected with the uses of those statistical procedures which until recently were the most formal and seemingly the farthest from mundane matters (and which are still such among those actually and extensively used). These are the procedures of significance testing and of setting confidence limits. Why does a behavioral scientist use a significance test? Or, better, why should he do so? The best answer is for purposes of communication.

This communication is sometimes between persons, and sometimes between roles within a single person. Indeed, Milton Friedman would distinguish them by assigning "calibration and communication" as a purpose of formal statistics. Here "calibration" means what I should have termed "adjustment of the investigator's optimism and pessimism". Perhaps it may better still be expressed as "aiding the investigator as data-gatherer and data-analyst to communicate effectively with himself as interpreter of appearances and assessor of theoretical importance".

This aspect is extremely important, possibly even as important, though I tend to doubt this, as the role of formal statistical procedures as means of communication *between* persons.

Each act of speaking or writing about one's results, formal or informal, is an act of communication, and its success depends on what is received, both as to extent and as to accuracy. Statements of significance or confidence should serve to improve communication; usually, and on balance, they do this. (Clearly this whole subject deserves deeper consideration than we can give it here.)

Communication has been studied and certain of its aspects quantified in modern information theory, which measures its amounts of information in bits, one bit being the maximum information provided by a choice between two alternatives. Clearly the settlement of the disjunction "significant" — "not significant" requires the transmission of one bit of information. (This is of course a very valuable bit, especially if the level of significance to be used is understood in advance. It is regrettable that we do not have a good measure of the *value* of information. Information theory certainly provides none such.) If we wish to know more about some investigator's result than merely the dichotomy of significant — not significant, we are likely to require several bits to specify what we have learned. To go beyond the level of a simple "yes" — "no" requires an increased effort, a greater channel capacity. And if it be true that information theory is relevant to mental habits, we should perhaps not be surprised to find many people who do not want to "clutter up their minds with any more bits of information," who consequently resist rather bitterly any tendency which might lead them to think in less black-and-white terms than "Smith's results was significant, but Jones's wasn't".

If we really must live with widespread commitment to such an attitude, we shall have to work out the best scheme we can, a scheme which will allow the use of "yes" — "no" alone and still manage somehow to allow us to get hold of what the data offer. But I estimate that the effort in preparing such a scheme, and the effort in using it, would not be worth while, that teaching people to think in terms of more than one bit at a time would require a far smaller investment. (It will rather clearly be desirable, in any event, to have a quite simple scheme, using some 3 bits, and a somewhat more complex one, using perhaps 10 bits, as intermediary techniques between "significance" — "nonsignificance" and a rather full assessment of the situation.)

Thus, while rigidity of separation into "significant" and "not significant" may possibly be necessary (but see G4 for some of its difficulties), it is possible that we can *all* learn to communicate more effectively about results, both with ourselves and with one another, using more flexible and useful codes.

C3. *DATA-GUIDED ANALYSIS*

Badmandments 23 and 91 refer to the relation between the data and what is done to it. At first glance they seem to contradict one another, one seeming to imply that the data should guide analysis, and the other that it should not. On careful examination, however, it appears that they do not contradict one another, but rather call for a combined policy, where *an* analysis is planned before seeing the data (preferably before gathering it) but the *actual* analysis is not *confined* to that which was planned in advance. Is this really the best way to proceed? What are the pros and cons?

Some would hold, indeed, that there is something unethical about allowing the data to guide its own analysis. Some of these would once have been (and some still are) purely mathematical statisticians, who sought exactness of probability statements and who saw no way to save this exactness if the mode of analysis was not prechosen. Others must have been urged on by feelings for which I have no ready analysis. The discomfort of the "purely mathematical" statisticians revolved mainly, in my judgment, around problems of multiple comparisons and complex experiments. These were, and seem to remain, the outstanding cases where the dangers of data guiding seemed to outweigh its advantages. Today there are available techniques, some of which will be alluded to in H3 below, which enable one, in both multiple comparisons and complex experiments, to allow the data to guide its analysis (within moderately broad limits), while preserving the same degree of exactness of the probability statements as would have been available if self-guided analysis had not been used. Thus a very large part of this objection has disappeared, and the manner of its disappearance has suggested ways which further development of new techniques may remove further parts. In the meantime, however, the principle that it is wrong for the data to guide its analysis has become an emotional commitment for too many. Even though its main reason for being has disappeared, we may expect this view to be clung to. But we need not join those who cling.

On the other side of the picture, it is even easier to argue that not letting the data guide its own analysis is unethical ··· not just statistically unethical, but scientifically unethical. If the data is really trying to tell us something, should we stop our ears to the answer, just because we didn't think of the question in advance? Clearly not if we are seeking knowledge. We cannot afford to seek knowledge at the price of maintaining no contact at all with the reality of the likely effects of random fluctuations, but since present-day statistical techniques (and even more those of the near future) allow us to combine increasing degrees of data-guidance of analysis with reasonable

control of exactness of probability statement, we dare not bind ourselves away from the data-guidance in the conduct of our analyses.

D. ORDERED CLASSIFICATIONS, THEIR CARE AND FEEDING

Perhaps the largest class of opportunities (whether good or evil) for analysis which are customarily ignored on a "do nothing, do nothing wrong, be not wrong" basis arise in connection with measurement. The subject of measurement has been discussed with much wisdom · · · and with much lack of it. It has been discussed just enough, and from sufficiently specialized points of view, to ensure that far too many people will "act scared", will refrain either from doing better what they already know how to do better or from inquiring into how they might learn to do better what they do. Many of us need to examine the reasons why we feel the way we do about measurement, and then ask if our feelings are at all justified.

We cannot deal with the subject exhaustively here, but we can try to illuminate some of its aspects. We deal first with ordered classifications and later (see E) with the choice of desirable modes of expression.

D1. WIDTHS OF CLASSES

Very many datums of behavioral science are expressed in terms of position along an ordered (linearly arranged) classification. Sometimes this classification is intrinsic, as when answers to a questionnaire are on a five-or seven-point scale ranging from "strongly disapprove" to "strongly approve". Sometimes this classification is observer-generated, as when families are placed as "working class", "middle class", etc. In either of these examples, and in many more, there is but little doubt either about the fact that the classification is ordered, or about what order is correct. (There are many instances of classification where this is not the case; they are not subjects for the present discussion.) How should we handle such information? How many classes shall we use initially? What penalties do we pay because of "misclassification?" Should we combine classes prior to further analyses? The answers to such questions are likely to depend on how we view the purposes, potentialities and perversions of ordered classifications.

We might begin by taking a rigid, logically seamless attitude toward classification, as Hempel does in his philosophical treatment (Hempel 1952), and require that classifications be definite, perfectly reproducible and without error. As a description of practical classification this is obviously quite unrealistic. (The distinction between males and females of *Homo sapiens* is probably as clear as any distinction of interest to the behavioral sciences, far clearer than most, yet the newspapers delight in telling us of occasional misclassifications.) In practice, classification is made with error.

The finer the classification, the narrower the classes, the more frequently will an independent reclassification *dis*agree with the original classification. But it is far from obvious that finer classes thereby produce less useful classifications. Classifying men into weight classes about 20 pounds wide will be more reproducible than classifying them into 2 pound classes, but there is little doubt that the latter classification provides more information. Indeed, classifying them into 16 times as many classes, into classes 2 ounces wide, would also provide more information than the classification into 2-pound classes, though communicating the additional detail may require more effort than the increase in precision justifies.

Now many will argue that all this discussion about human weight is true but irrelevant, for this situation differs from that common in the behavioral sciences in two ways: first, a continuous scale of weight underlies the classification into weight groups, and second, perhaps even more importantly, the measurement of weight is a *physical* measurement conducted on a scale of very prestigious (and in truth very desirable) properties · · · a ratio scale. To me such arguments seem very weak, once examined.

Suppose that no physicist had ever lived, and that the only way of comparing "weights" was by a two-pan balance (without a scale) and a set of weights. Suppose further that no one had ever thought of putting two weights in a single pan. What could international standardization have done? It seems to me that its effort would have been devoted to the preparation of standard bodies, numbered in some way, and assembled in sets, each individual of a set differently numbered and different individuals of the same number chosen so as to very, very nearly balance one another. Then weighing, of a man or of a bag of potatoes, could be conducted by comparing the unknown with each standard body of some standard set, thereby assigning the interval between standard bodies in which it fell.

If this were the case, it would still be true that, although no trace of a ratio scale would be available, weight comparisons of men with a closely spaced set of standards, would be more informative than

comparison with a very coarsely spaced set. Thus the ratio-scale aspect had nothing to do with whether we learned more or less from a finer classification.

The belief that a continuous scale makes an important difference is similarly not sound. The discussion we have just given would have gone smoothly, if not continuously, along the same course if the weights of all objects concerned, humans or standard bodies, were always exact multiples of one pound, or of one ounce, or of one grain, so that every weight was expressible on a discrete scale · · · so that the continuous scale had no physical reality.

There seems to me no escape from the conclusion that, so long as the class boundaries are well defined *in some average sense,* we learn more from a finer classification than from a coarser one, even though we expect poorer reproducibility for the finer classification, at least *as measured by fraction of agreements on independent reclassification.*

Now it could be argued that there are many behavioral science situations where the boundaries would become much less well-defined if there were an attempt to use finer classes. Doubtless there are such circumstances. It is clearly a subject-matter question how often this happens, one where actual inquiry is better than expert judgment (behavioral science experts, that is), which is better than an outsider's impressions (such as mine). But each of us is entitled to his own opinion. My opinion, strengthened by listening to such remarks as "we wanted to divide them further, into 'upper middle class' and 'lower middle class', and into 'upper working class' and 'lower working class', but when we tried it the numbers were too small", is that there are relatively few such instances.

D2. DIVIDING THE MIDDLE CLASS

Let us examine this last instance more carefully. How could we lose information by dividing both "middle class" and "working class" into "upper" and "lower"? It seems most unlikely that such a refinement would affect our decision, in any but an exceedingly small fraction of all cases, as to whether a family was "working class" or "middle class". (And I doubt whether the changes that did occur would, on the average, increase *mis*classification.) Similarly, I cannot believe that those classified "upper working class" would, as a group, be actually lower in the class structure than those classified as "lower working class". The worst one could conceive, and this is very nearly too hard for me, is that families might be randomly assigned to the upper and lower segments of the "working class". As is set forth in

more detail in Appendix S4, random splitting is only as much worse than no splitting as perfect splitting is better, while rather poor splitting is still an improvement. Consequently, if we have any real basis at all for splitting, it is relatively certain that we shall be better off to split.

Now it is likely to be argued that, if we halve, or further divide, the classes we have done an evil thing, because more frequent "errors" in classification will blur the meaningfulness of the classes. Let us examine halving the middle class from this point of view. It is indeed true that the fraction of all families actually classified "upper middle class" about whom we are uncertain as to whether they really belong "in" is greater than the corresponding fraction for the middle class as a whole. But so what? It is equally clear that the group of families classified "upper middle class" is a more homogeneous group than that including all those classified "middle class". As a group, and it is to the resulting group that further analysis will apply (not to group boundaries), there is less blurring for the smaller group. And this is so because added variation *within* the larger group contributed to blurring of the group image in just the same way as, and usually to a greater extent than, difficulties with assignment to classes can contribute via misclassification.

If the smaller classes are less blurred, how then could a prejudice develop against them? Two reasons seem most obvious:

(1) statistical techniques of the required flexibility did not seem to be available;

(2) certain sorts of misinterpretation were possible for the naive.

We shall return at a number of places below to the question of how currently available statistical techniques can be used to deal effectively with finer classifications. It is easy to understand reason (1) being once strongly felt, but today it no longer offers adequate ground for a prejudice.

Reason (2) deserves more discussion. The simplest sort of misconception involved arises where subdivision into "upper" and "lower" has limited effectiveness, and where, as a consequence, the average apparent difference between the upper and lower segments of a class is much less than it would have been for a "good" sub-classification. The incautious investigator might then conclude that the step between classes was much larger in comparison with the gradation within classes than was in fact the case. If fine classes are used, their users must be prepared to recall at frequent intervals that their establishment is fallible. They must recognize that they are "living dangerously". This is uncomfortable, since we all like overall feelings

of surety and confidence. But in assessing such discomfort, we *must* remember that we have actually learned more from the finer classification; our only danger is from believing we learned too much more.

There is a Scots proverb, quoted by John Buchan somewhere, to the effect that "A man may have a gey fine hoose, but he maun sit loose to it". This has many and stringent applications to the analyst and interpreter of data, who may indeed have "a gey fine body of data" but who, if he is going to get the most out of it, must "sit loose" *both* to it *and* to many of the interpretations he bases on it.

Perhaps an example from quantitative measurement will illustrate the point. The teachers of primary and secondary arithmetic are likely to purvey the doctrine that if you are not sure of a figure, you drop it. At a higher level of sophistication, surveyors and navigators (to whose arts not all behavioral scientists have been exposed) are likely to carry one or two extra figures through the computation (and then cut down somewhat at the end). At a higher level of sophistication, or so one might suppose, should come the makers of mathematical tables, who have traditionally been the purest of the pure. Though they are not studying the empirical world, they face a similar problem, because their numerical calculations (not conducted in integers or rational fractions) are made with limited accuracy. If calculations to a particular accuracy yield .1349 when more precision would have yielded .1359, and the answer is desired to only 2 decimal places, then .13 will be entered in the table, when .14 would have been closer. Instead of being correct to within .0050, such a table is at best correct to .0059. Some table-makers would take .0059 but boggle at .0061. Others would draw the line between .0051 and .0052. Indeed, a few purists, knowing that a true value was between .26498 and .26502 would refuse to enter .26 since it might be off by .00502, which they regard as too great an error for a two-place table. (Since errors up to .00500 are inevitable in two-place tables, the economics of these judgments are far from obviously sound, at least to a statistician.)

With all these diverse views at hand, what are we to do with quantitative measurements such as weights, heights, voltage, etc., some of which are made in duplicate? Suppose that values are originally written down to enough figures so that duplicates disagree very frequently. How much should we round off? The idea is abroad that we should round off until most duplicates agree. This idea is statistically unwarranted. Once we have cut back the recorded precision till as many as 10% or 20% of the duplicate pairs are identical, we have reached the point where further cutting back may discard detectable

amounts of information (Appendix S; also Tukey 1950). In the marginal situation a substantial % of duplicates will differ by 4 or more steps.

If we were to use the same standard on classifications, as well we might, provided (i) the juice-extracting power of the statistical techniques used upon classifications were equally great as that of those used on quantitative measures (this is nearly attainable), and (ii) the additional costs of computation were not significant, then we would not consider a set of classes too narrow unless classifiers could not agree precisely in *less than* 10 or 20 percent of instances classified. Compared to today's usual practice, such a rule would result in very narrow classes indeed. Granted that such a standard is likely to be too stringent, not only because of the provisos above, but also because uncertainties of order would be likely to arise for the kinds of extremely narrow class which would have to be defined, it is still true that ***much more is to be gained from finer classes than from coarse ones.***

D3. INADEQUACY OF BROAD CLASSES

One of the main reasons for introducing broad classes for some variable is so that the effects of that variable may be "controlled". Thus we may be interested in the descendants of two groups of immigrants, one from Atlantis and the other from Mu, and we may wish to compare their incomes, controlling "of course" for social class. Let us suppose (i) that social class is really a continuous variable, even if we may not know how to measure it on a continuous scale, (ii) that average income varies linearly with position along this scale, in ***exactly*** the same way for both groups of descendants, and (iii) that, for each group, social status is normally distributed, the two groups having the same variance. Then the figures in Table 7 are perfectly possible, and perfectly consistent with these hypotheses. It would not be unnatural for the report of such an investigation to read "even after controlling the effects of social class, average income of those of Muan descent substantially exceeded those of Atlantean descent; this difference is probably to be ascribed to · · · extended family · · · facility toward accepting industrial society · · · strong motivations · · · ." Such language could of course be completely wrong, as the example shows. Such language is thus always nearly completely misleading, since the effects found could well be the result of incomplete fineness of classification. Table 7 illustrates this in detail. While Table 7 shows that the seven-class breakdown did a much more thorough job of "controlling" social status than the three-class breakdown, its success was far from complete.

Table 7

Hypothetical example showing failure of broad classifications to "control" the effect of a variable* when comparing "A" with "M". (Two versions.)

Social Class	Distribution** A	Distribution** M	Average annual income ($) A	Average annual income ($) M	Average annual income ($) (M-A)
1	—	1.0%	—	(4078)	—
2	0.3%	10.0%	(3506)	3558	—
3	7.8%	43.0%	2902	2993	+91
4	40.9%	39.0%	2304	2414	+110
5	35.0%	6.4%	1778	1835	+57
6	11.0%	0.5%	1363	(1382)	—
7	5.0%	0.1%	968	(1039)	—

Social Class	Distribution** A	Distribution** M	Average Annual Income ($) A	Average Annual Income ($) M	Average Annual Income ($) Diff
1	—	1.0%	—	4078	—
2-4	49.0%	92.0%	2407	2809	+402
5-7	51.0%	7.0%	1609	1792	+183

*Mean annual income for infinitely narrow social classes the same for A as for M.

**Rounded values. Values to 1 more decimal are for A: 0.32%, 7.76%, 40.92%, 35.01%, 11.04%, 4.95%; for M 1.00%, 10.2%, 42.96%, 39.01%, 6.40%, 0.53%, 0.08%.

How can it come about that the use of broad classes is inadequate to control the effect of an extraneous variable? Figure 1 shows how this can happen. When we separate out all the cases which fall in the broad class, we obtain a distribution of the continuous variable that is confined between given limits. True, but the shape of this confined (truncated, censored) distribution is *not* given. As Figure 1 shows, this shape depends upon *where* the distribution of the underlying continuous variable peaks up.

Now, at least in the simplest situations, what matters is the ***mean of the underlying continuous variable*** for all cases falling in the broad class. These means, for Group A and Group B, are the centers of gravity of the shaded areas in Figure 1. Clearly, these centers of gravity need not be the same. Consequently, the groups picked out as belonging to the given broad class ***need not be comparable*** in terms of the continuous variate.

This example is ***not*** intended to convince the reader that "controlling" variables in broad groups is useless or unwise. It would be a serious mistake to come to any such conclusion, since such "controlling" is effective, useful, and indispensable. The purpose of, and the only appropriate lesson to be learned from, this example is that such "control" is far from complete; that its incompleteness can have noticeable and apparently significant effects, that control into finer classifications can be more effective. (For a possible technique for improving the use of broad groups see Appendix X.)

D4. HOW SHALL WE SCALE THE RESULTS?

The first roadblock in the way of applying sensitive techniques of analysis is the investigator's reluctance to assign numerical values to each class when he faces a classification that entails more than two classes. This form of mental paralysis appears to be an anaphylactic reaction to successive injections with statements about the importance of measurement on proper (not just appropriate) scales. (That anaphylaxis has taken place is obvious from the tremendous extent by which the reaction exceeds that appropriate to the situation.)

What are the facts? Let us suppose that we have five classes, duly arranged in a reliable order. Let there be ***ideal*** scale values to attach to these classes, values which we do not know, but which certainly increase as we go up the order from one class to another. How weird might these ideal scores be? Let us suggest some possibilities, bearing in mind that we shall lose nothing of importance by fixing the score of the lowest class at 0, and that of the highest class at 10. Certainly some of the following seven possibilities are rather extreme. If we assess our proposed actions in terms of how satisfactorily they will behave in the face of each of these possibilities, we should be able to learn a considerable amount about what the effects of choosing different actions will be.

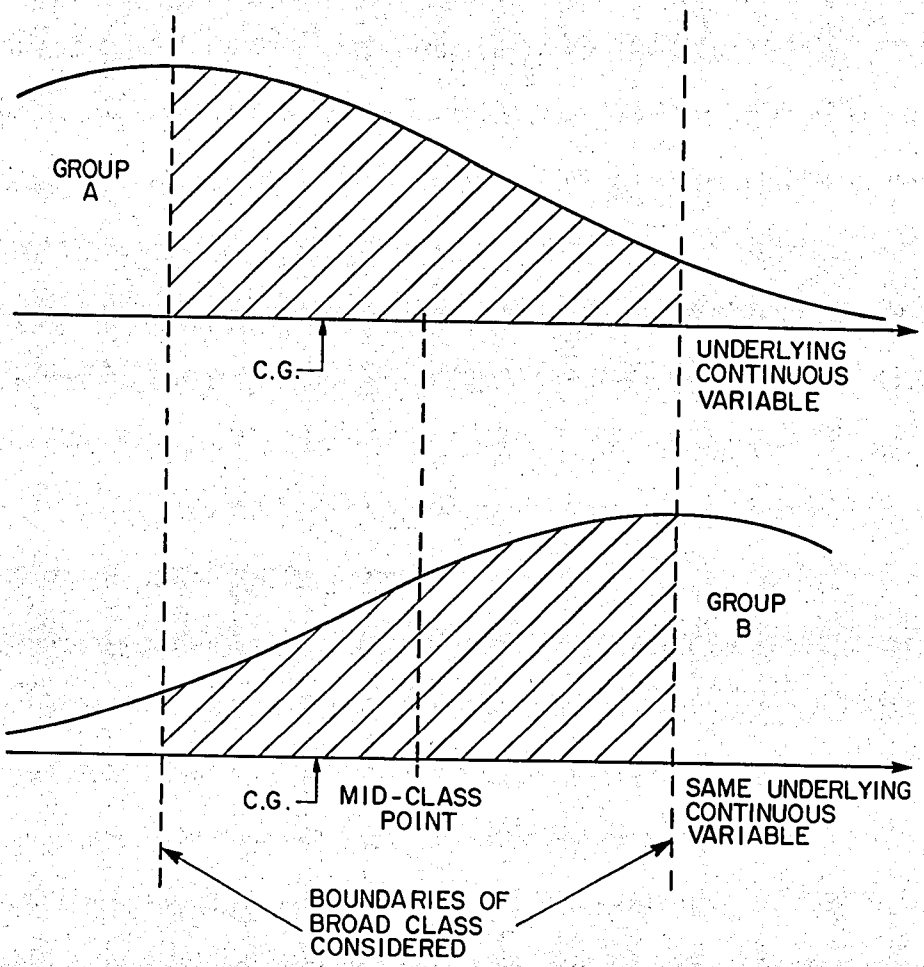


Figure 1. Underlying continuous variable replaced by a broad class. Dependence on center of gravity (C.G.) of continuous variate for broad class on location of distribution

A	*B*	*C*	*D*	*E*	*F*	*G*
0	0	0	0	0	0	0
1	1	2	5	1	2	3
2	2	7	6	3	5	6
3	9	8	7	6	8	8
10	10	10	10	10	10	10

The "party line" (of the party of mental paralysis) is that we are *safe* if we consider only breakdowns into two classes, for then we can choose both the numerical values without loss of generality and, therefore, we should combine our five classes into two. Which two? The party line sayeth not, and it is reasonable to assume that any way of dividing the five among an upper and lower group is entirely acceptable. But just what is the consequence of such a reduction in the original number of classes? It is just that certain classes are scored with one value, which without loss of generality we may take as 0, and all others with another, which we may equally well take as 10.

The "party line" disapproves most strongly of merely assigning equally spaced values, which here would be 0, 2.5, 5, 7.5 and 10, to the classes. This is evil · · · because we don't know that it is the ideal thing to do! But what are the actual consequences? We clearly need some overall measure of agreement and disagreement between the different scalings. If we knew the exact frequency with which the various classes occurred, it would be natural to calculate (the square of) the correlation coefficient (over individuals) between items. In the absence of such detailed information, it is natural to treat each class as if it were equally frequent, and to calculate *formal* correlation coefficients between pairs of scalings. (In general, assumed *equal* frequencies will tend to make formal correlation coefficients fall somewhere near their lowest possible values. This proves further justification for this choice.) Table 8 presents the squares of such formal correlation coefficients between each of the 5 alternative scorings and each of the 7 suggested *ideal* scorings.

The natural and fair comparisons are between any single linear scoring and a random choice among the dichotomies. *For any of the ideal scorings* considered in the table, the average performance of the four dichotomies *never comes close* to the performance of linear scoring. The chance that a randomly chosen dichotomy will do better than the linear scoring is never more than 25% for any of these seven ideal scorings, and is usually zero. (The doubting reader is encouraged to repeat the calculations for his own assumed scorings. However, Robert Abelson and I have been looking deeper into such matters. It is easy to

Table 8

Quality of approximation of various dichotomies, as compared with linear scoring, for seven possible ideal scorings. (Measured in terms of squared correlation coefficients.)

Ideal Scoring	Dichotomies 1/2	2/3	3/4	4/5	aver.	best	worst	Linear score	Exceed-ances*
A	.20	.35	.58	.92	.51	.92	.20	.70	1/4
B	.27	.48	.97	.44	.54	.97	.44	.94	1/4
C	.51	.90	.61	.37	.60	.90	.37	.95	0/4
D	.68	.56	.59	.42	.54	.68	.42	.84	0/4
E	.30	.62	.81	.68	.60	.81	.30	.95	0/4
F	.46	.79	.79	.46	.62	.79	.46	.99	0/4
G	.57	.79	.67	.41	.61	.79	.41	.97	0/4

*Number of dichotomies doing better than the linear scoring, expressed as a fraction.

"Ideal" scorings *A*	*B*	*C*	*D*	*E*	*F*	*G*	Assumed scorings 1/2	2/3	3/4	4/5	Lin.
0	0	0	0	0	0	0	0	0	0	0	0
1	1	2	5	1	2	3	10	0	0	0	2.5
2	2	7	6	3	5	6	10	10	0	0	5
3	9	8	7	6	8	8	10	10	10	0	7.5
10	10	10	10	10	10	10	10	10	10	10	10

show, for this case of 5 groups, (i) that if the ideal score is ordered in the same way as the classes, there cannot be more than one chance in four that a random dichotomy does better than the linear scoring, (ii) whatever be the ideal scoring, the average unsquared correlation coefficient for the dichotomies is less than 4/5 of that for the linear scoring. Choosing and calculating other examples cannot alter the picture substantially.)

Having been forced to abandon an unspecified dichotomy, the last-ditch defenders of the party line will presumably fall back on comparing the best dichotomy with linear scoring. Quantitatively, they can only claim a case for the two most extreme of the seven possible proper scorings considered. This is already not much help. But worse

is to come. The logic of any such position is nonexistent. If it is right for the protagonists of dichotomies to use enough insight into the problem to allow them to pick the best of the four dichotomies each time, then it is hard to see why those who (in the absence of insight) favor linear scoring are not equally entitled to use the same insight to choose a more effective modification of linear scoring. When they do this, they are almost certain to be ahead of the dichotomizers.

The pragmatic conclusions are, I believe, completely clear. If you must not use insight, use linear scoring rather than dichotomizing. If you may use insight, and have a reasonable amount to use, use it to modify the linear scoring, not just to choose a better dichotomy. In terms of getting the most out of the data, dichotomizing is dangerous and wasteful.

D5. THE MEASUREMENT OF CLASSIFICATION QUALITY

Ian Campbell Ross has pointed out to me that, since it is customary to publish evidence of the reliability of one's classifications, any proposal for the use of sensibly narrow classifications is unlikely to be widely accepted unless it is accompanied by a suitable way of measuring classification quality. What choices have we to consider? Those who labor with tests and measurements use reliability measures based upon test-retest correlations. If we are prepared to impute a numerical scale to our classes, we can easily use a suitable modification as a measure of classification-reclassification reliability.

It is natural to seek for a simpler index; perhaps to try to say that, while we can admit classes so narrow that independent classification will move many individuals into an adjacent class, we dare not use classes so narrow as to have any appreciable fractions of reclassifications that result from moves by, say, two or more classes. Such a view would be doubly wrong, wrong both in detailed fact and in principle. As shown in S5, below, efficient classes will be so narrow that a substantial fraction of reclassifications will be shifts of two or more classes. Moreover, what is far more crucial, using any such criterion would be a judgment on a false basis.

The questions:

(1) are the classes narrow enough to make efficient use of classifying ability; and

(2) is our classifying ability great enough to make classification useful in this particular problem

are quite distinct and separate questions. The answer to either may be "yes" when the other is "no." Thus a policeman's-eye estimate of a suspect's weight is surely efficiently utilized when given to the nearest pound, but is of no practical use in discriminating among a group of teenagers whose spread in weight is only 4 pounds. And the three-point scales used by the Gluecks (Glueck and Glueck 1950, pp. 68ff.) are surely strong enough to make classification useful in their problem, and equally surely far too coarse to make effective use of the classifying ability at their disposal.

To answer the question as to whether a classification has enough power to be useful, we should make use of some reliability measure. If we must have a particular standard method for general use in a wide variety of circumstances, we must select a way of assigning scale values to the various classes. Especially so long as we are concerned with reliability only, the center of gravity of the corresponding area under the standard normal distribution seems quite reasonable. Leverett's table (Leverett 1947) can be used without interpolation (integer %'s being quite close enough) and without accepting any specific views about a "true situation."

A particular example is carried through in Table 9 as an illustration. (Keeping two decimals in the answer is surely informative enough; one decimal may suffice in many instances.) It is important to emphasize that this is one of many indices which might be used for this purpose. The vast majority of these indices would work satisfactorily. And there is no clear theoretical reason for preferring one to another. The great reason for the choice of the index illustrated in Table 9 is its case of calculation.

To answer the question as to whether a classification has enough classes to make efficient use of the classifying power, we need an appropriate indication of what that power really is. Reclassification by the same judge at another time is not likely to be completely independent reclassification. Independent reclassification by judges, both of whom belong to a group of judges used to cross-checking one another, will give closer agreement than independent reclassification by judges who have only read the instructions and criteria. And so on. Clearly there is a place for considerable wisdom in determining what sort of reclassification fairly indicates the relevant kind of classifying power. (The existence of various kinds of reliability is familiar to all who measure reliability in mournful numbers.)

A broader gap is of serious importance in many circumstances. An infinitely detailed book of infinitely detailed rules can produce near perfect classification, but the infinitely small details of the classification

Table 9

Example of the calculation of an index of reclassification consistency or reliability using Leverett's 1947 table.

(Observed distribution of classification and reclassification)

	A	B	C	D	Total
A	54	12	4	2	72
B	13	65	3	0	81
C	5	3	59	12	79
D	1	0	6	46	53
Total	73	80	72	60	285

(Calculation of scores for each class) mean of two class n's*

	no.	no.	%	%	Score (Leverett 1947)
A	72.5			25	1.27
		212.5	75		
B	80.5			29	−.27
		132	46		
C	75.5			26	−.45
		56.5	20		
D	56.5			20	−1.40

(Calculation of index itself)

$$\text{index of reliability} = \frac{54(1.27)^2 + 12(1.27)(0.27) + \cdots + 6(-.45)(-1.40) + 46(-1.40)^2}{72.5(1.27)^2 + 80.5(0.27)^2 + 75.5(-.45)^2 + 56.5(-1.40)^2}$$

$$= \frac{87.22(1.27) + 32.71(.27) - 36.19(-.45) - 65.84(-1.40)}{92.075(1.27) + 21.735(.27) - 33.975(-.45) - 89.1(-1.40)}$$

$$= \frac{218.0628}{260.23245} = 0.83 .$$

* Mean of classification and reclassification counts.
(Note: %'s are taken to nearest whole %.)

are almost certain to fail to reflect what is supposed to control the classification. Two individuals, whose "true" scale locations are exactly the same, may, for example, be classified consistently and repeatedly into widely separated classes. In such situations, the adequacy of class fineness should not be judged in terms of agreement of classification and independent reclassification of the same individual, but rather in terms of agreement of classification of "truly equivalent" individuals. Direct evidence about this latter sort of agreement will often be too hard (if not impossible) to obtain, and it may be appropriate to choose class widths on the basis of a subject-matter expert's belief that the agreement of classification of equivalent cases is notably less than the agreement of classification and reclassification.

This sort of judgment can be reasonably good evidence that it is not worthwhile to go to smaller classes, while, as is abundantly documented in S5 below, an observed agreement of only 50% for independent reclassification into the more finely divided classes is evidence not that the classification is too fine but, rather, that finer division may well be quite useful.

D6. THE CONNECTION BETWEEN SCALE TYPES AND STATISTICS

We have assigned scale values to ordered classifications in a way that some would judge to be blithe and arbitrary. So long as we only look at the resulting numbers there will be little conflict. But when we come to combine and dissect them, to analyze them in as wide a variety of manners as seems to prove useful, then there will be objection. For some will have read S. S. Steven's discourses on how each individual statistical procedure, more specifically each individual summary statistic, should only be used on data of a suitably high scale type (Stevens: 1946, 1951, 1955, 1959). As Luce (1959, p. 84) summarizes the matter:

> "· · · limitations that the scale type places upon the statistics one may sensibly employ. If the interpretation of a particular statistic or statistical test is altered when admissible scale transformations are applied, then our substantive conclusions will depend on which arbitrary representation we have used in making our calculations. Most scientists, when they understand the problem, feel that they should shun such statistics and rely only upon those that exhibit the appropriate invariances for the scale type at hand. Both the geometric and the arithmetic means are legitimate in this sense for ratio scales (unit arbitrary), only the latter is legitimate for interval scales (unit and zero arbitrary), and neither for ordinal scales. For fuller discussions, see Stevens: 1946, 1951, 1955; for a somewhat less strict interpretation of the conclusions, see Mosteller, 1958."

The view thus summarized is a dangerous one. If generally adopted it would not only lead to inefficient analysis of data, but it would also lead to failure to give any answer at all to questions whose answers are perfectly good, though slightly approximate. All this loss for essentially no gain. (We return at the end of the next section to an analysis of why this seemingly logical argument can be so misleading.)

More precisely stated, the limitations discussed by Luce do not control which statistics may "sensibly" be used, but only which ones may "puristically" be used. Consider an individual physical or chemical measuring device of any specific sort. It will have systematic errors of calibration, like all individual measuring devices, which will depend upon the part of its scale in which a measurement falls. Suppose further that, as is so frequently the case, these systematic errors are modest, rather than minute, and vary systematically but slowly with location on the scale, and that the results of using this device are to be the subject of statistical analysis. What measures of typical value and spread dare we use?

The measurements are not on an interval scale, in Steven's sense. For the results of another individual measuring device, separately calibrated and making quite different systematic errors, would have the same quality and validity as those obtained with this particular instrument. And the relation between these two equivalent scales is not of the form $z = \alpha + \beta y$. In Steven's eyes, at least as interpreted by others, a scale that is not an interval scale is only an ordinal scale. To such eyes it is only sensible in such a case to use those statistics which are invariant, or better, covariant, under all monotone increasing transformations.

If we have a sample of 10 such measurements, this principle would forbid us to calculate the mean of all 10, or the mean of the central 6, etc., because means, truncated means, and the like, are *not* covariant under all monotone transformations. Stevens would, of course, allow us to use the median of the sample. To use the median may be to lose a noticeable amount of efficiency, but one at least gains some advantages in return. Comparing the location of two populations in terms of the medians of two corresponding samples is not impractical, may indeed often be advantageous, though it may also be wasteful.

What if we want to compare the spreads of two populations in terms of the two corresponding samples? It is natural to compute a measure of spread for each sample, and then to compare them. There appears to be a wide choice of measures of spread, including:

the standard deviation of an entire sample;

the standard deviation of a truncated (censored) sample;

the range of an entire sample;

the range of a truncated (censored) sample;

the interquartile deviation of the entire sample.

Surely most of these will serve, even if some may not. Not in Steven's eyes. None of these measures of spread is covariant under all monotone increasing transformations; indeed it is easy to see that *no measure of spread* is so covariant. Thus the Stevens view leads to abandoning the question as to whether two populations have the same spread.

If our measurements were on a scale of which the most that could be rightly said was that it was defined only up to a monotone increasing transformation, that it was indeed merely ordinal, then this conclusion would be quite correct. To compare the spreads of two populations measured on a merely ordinal scale is senseless if the two populations do not have a very substantial overlap. The question must be "If a scale is not an interval scale, must it be merely ordinal?".

D7. THE MEASUREMENT OF TEMPERATURE

Let us turn to the history of physics, to the days before the development of the thermodynamic scale of temperature. How were temperatures measured? With one of any of several kinds of thermometers. (In the early days these would have been liquid-in-glass thermometers with different liquids enclosed in different kinds of glass. In later days they would have been gas thermometers using different gasses at different densities.) Would there be agreement between the different kinds of thermometers? Approximate agreement, yes; exact agreement, certainly not. Would any one kind have sufficient theoretical support to be chosen as *the* standard over all others? No. Clearly temperature was not measured on an interval scale in those days. But equally clearly, it made good sense to compare the spreads of two populations of measured temperatures, and to calculate the arithmetic mean of a group of temperatures. Temperature was not measured on a *mere* ordinal scale. It was measured on a scale which, though not an interval scale, was still quite well defined.

Temperature in those days is a clear example. Today a wide variety of other measurements are less clear examples. Not every quantity measured on an ordinal scale that is not an interval scale is such as to deserve the calculation of a sample mean or a measure of sample spread. But there are many that do deserve treatment of such quality, and it would be wasteful not to take advantage of the opportunity to learn more about many things by making such calculations.

As described by Luce, the Stevens position seems cogent and logical. Yet we have indicated how it fails. What are the reasons for its failure? The two most fundamental seem to stem from:

(1) A lack of adequate recognition that knowledge is approximate, not precise.

(2) A lack of appreciation that all useful conclusions are not fundamental.

From (1) comes the failure to recognize that many scales, such as the early scales of temperature, are *approximate* interval scales. Almost all liquid-in-glass thermometers show general agreement as to a temperature scale. Large differences in variability between two populations of temperatures remain large on all scales. And only large differences can be detected reliably with samples of reasonable size. Here the approximation to an interval scale was close.

Many assignments of scale values to ordered classifications, assignments which may be either equally spaced or more carefully or appropriately chosen, produce approximate interval scales, where the approximation is much rougher than for old-time temperature. But the approximation is still there; arithmetic means and measures of spread can be very useful, provided they are interpreted with proper caution. It is here that (2) enters. If a finding that "variance increases as we move up the scale" is only useful if it can be taken as a contribution to the fundamentals of psychology, then we must be very careful about making such statements. But if it serves to guide us, perhaps in the design of an experiment, perhaps in the choice of a method of statistical analysis, perhaps in the directions in which we seek new or modified theories whose confirmation we realize must rest on approximate results, such a statement, although resting on a wholly approximate foundation, may be very useful.

One reason for the feelings of those who believe that precise scale type should limit the use of statistics may well be the practice, entered into by too many, of regarding statistical procedures as a sanctification and a final stamp of approval. Results based on approximate foundations must be used with the underlying approximation in mind. Those who seek certainty rather than truth will try to avoid this fact. But what knowledge is not ultimately based on some approximation? And what progress has been made, except with the use of such knowledge?

If a crudely assigned scale, perhaps followed by a handy transformation, leads to data which fits nicely into one of the additive patterns associated with the analysis of variance, yielding only very small interactions, then an empirical fact has been discovered. Arithmetic means and measures of spread will have been calculated from values of which it could not be confidently asserted in advance that they deserved such treatment. But the results will have shown, by their clear additive behavior, that they did deserve it.

An oversimplified and overpurified view of what measurements are like cannot be allowed to dictate how data is to be analyzed. In particular, it *may* be reasonable to apply relatively sophisticated analyses to equally spaced values (or more carefully chosen standard scalings) which have been "arbitrarily" assigned to an ordered classification.

E. MODES OF EXPRESSION

The title of this part was chosen advisedly with the intention of avoiding "loaded" words. The idea it is intended to convey, an idea which appears to me to be correct, is that we have decided what aspects of which portions of the data we wish to express numerically, and we have now to choose a mode of numerical expression, one which will be most useful to us for our purposes of analysis. We are, at this point, trying to tune our ears to hear what the data are trying to say to us. Good data try, much harder than most of us realize, to tell us what is going on. We need receptive ears, prepared to hear Scriabin when we expected Scarlatti, but *not* insisting that what we hear must be either.

In this tuning process, graphical techniques can be of great service, especially when we draw alternative crude graphs to help us listen flexibly, rather than single, definitive, professional graphs, such as a "deaf-to-data" investigator might plan before seeing the data. We shall give but little attention to this graphical aspect of analysis here, only a small fraction of the amount it deserves.

But what are we really doing when we plot and replot the same data on various kinds of graph paper with differently spaced rulings? We are experimenting with different modes of expression for the two variables represented along the axes. When will we be likely to feel the happiest? Probably when we find scales such that the "curves" are straight lines. When should we feel happiest? Probably when we find scales such that the "curves" for y against x for two different portions of the data (which may correspond to two countries, two occupations, two education levels, etc.) are *parallel* straight lines. For in this latter case the description of what we have found is surely as simple as possible. Our response, measured vertically, increases in fixed proportion to increases in our explanatory variable, which is measured horizontally, *and* the differences between portions (countries, occupations, educations, etc.) are described by a single number, the vertical distance between the curves. (The simplicity of the graphical picture is reflected in the simplicity of the numerical description: one

number for the (common) slope of the lines and one for the shift between them.)

Described in such terms, such an ideal state sounds simple, so simple as to be unuseful, unconnected with the higher principles of "proper" measurement, even perhaps simple-minded. We shall see whether this is so.

E1. MONASTIC MEASUREMENT

Just as some have done for mathematics, measurement may be divided into "monastic" and "secular". The analogy of the "high church" view, which we naturally call the "high monastery" view, is surely that provided by Norman R. Campbell, whose two books (Campbell 1920 and reissued 1958, 1928) have been the source, proximate or remote, of many fears that assignment of numbers, many of which would have been perfectly useful, were not "measurements". These books contain many deep insights, both into classical physics as a science, and into measurement as classical physics practiced it, and still practices it. Their intellectual authority is obvious, their message is clearly meaningful; we must beware only to be sure that we are affected and guided by the true substance of Campbell's inquiries rather than by superficial considerations (among which may be included some of Campbell's own views and statements).

We need here only to be concerned with his idea of fundamental measurement, which hangs upon the twin hooks of comparison and concatenation. He assumes of some characteristic of "objects" that:

(1) we may determine for each pair of "objects", A and B, whether $A > B$, $A = B$ or $A < B$;

(2) we may "combine" each pair of "objects" to form a new object, $A + B$; and

(3) that these comparisons and concatenations are intra- and inter-related in suitably axiomatized ways.

Leaving aside Campbell's treatment of error, which leaves me (I believe both as a statistician and as an ex-physical scientist) quite unsatisfied, there is little doubt that measurement which fulfills Campbell's requirements, exactly or approximately, is measurement which deserves the highest social status, the highest prestige that we can today imagine.

Certain remarks are of importance at this point. Campbell says (1928, pp. 41, 42):

> "Now the devising of methods for judging > is the chief problem of the experimental art. For the range of possible methods is enormous; any effect, however remote and indirect, of a change in a magnitude [i.e., in the characteristic being measured jwt] provides a possible method; and the most ingenious experimenter is he who can see ways of using very remote effects."

In terms of explanatory variables and responses, this translates into: "So long as an increase in the explanatory variable increases the response, that response is a proper candidate for use in measuring the explanatory variable: the best candidate will be the most sensitive response." In making this translation, we have made a analogy between Campbellian measurement and explanation of response which is not the one that is likely to seem most natural, and we must explain both the analogy and our choice of it.

Suppose that we are studying the combined effects of education, reference groups, work groups, and family groups on political opinions. The more or less normal approach would seem to run as follows:

(1) first let us decide how to measure political opinions, (after all this is what we are studying);

(2) then let us try to describe how political opinions (thus measured) vary with the factors with which we are concerned.

Is this approach sound? It may appear so now, but let us study its soundness by setting up a physical analogy. We shall soon wince.

E2. QUADRUPLET WEIGHING

Let it be supposed that we have brass weights, gold-plated weights, aluminum weights and quartz weights. Let it further be supposed that we wish to study the weights of combinations made up of one brass weight, one gold-plated weight, one aluminum weight and one quartz weight (strictly analogous to one kind of education, one reference group, one work group, one family group). Let us suppose that one member of our research team has seen the type of letter scale (very ingenious, indeed) where a plate hangs from a pivot in the plate, a suspension link that is also a pointer, as in Figure 2. As heavier objects are attached to such a scale, the plate balances in different positions, and

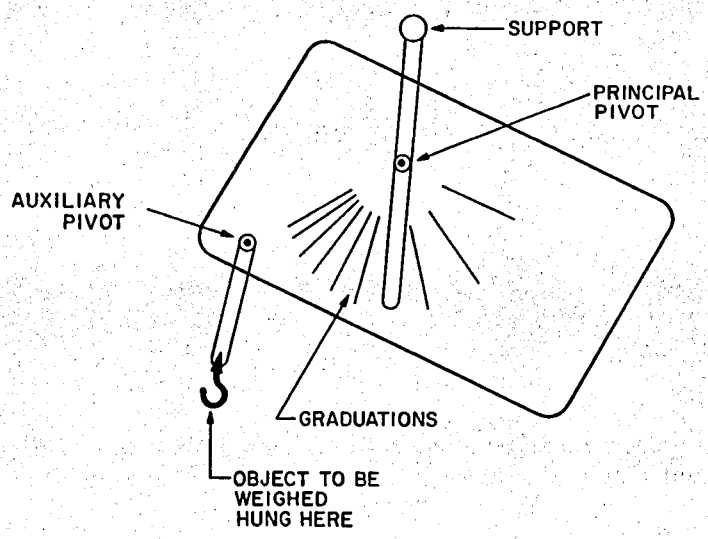


Figure 2. Simple Letter Scale

the graduations on the plate move past the pointer. To weigh in ordinary terms, as in grams, ounces or pounds, these graduations must be unequally spaced. Let us suppose further that this member of the research team then builds a weighing device of this sort of the best workmanship, employing kinematic design, jewelled contacts, and the best knife edges, *and* that he provides a *uniformly* calibrated scale over which the pointer is to travel.

What now happens when we start weighing our quadruples (one weight each: brass, gold-plated, aluminum, quartz)? We get numbers, highly precise numbers. And these numbers respond in reasonable ways when we exchange one quartz weight for another quartz weight; they always change in the right direction. But there is most serious interaction! Changing from quartz weight Q_1 to quartz weight Q_2 has a different numerical effect when brass weight B_1, gold-plated weight G_1, and aluminum weight A_1 are present than when brass weight B_2, gold-plated weight G_2 and aluminum weight A_2 are present. What should we do?

By following the analogy of what might be considered the standard approach of factorial experiment, we find ourselves in a complex and troubling situation. And we secretly know that we could have avoided most of this difficulty by choosing a better scale of "weight" to begin with. It would be desirable to avoid such situations in every instance. But we usually (or perhaps only often) lack the secret

knowledge which could have saved us in this special instance. In instances resembling the influences-on-political-opinion example we are almost sure to lack such secret knowledge when we begin. Thus we must be prepared to encounter such troublesome and complex situations. We must study the possible ways out of difficulty, and be prepared to choose and use one or more of them in many situations.

E3. WAYS OUT

When working with weight quadruples in this way, there may be three alternatives:

(1) We may choose to represent our arbitrarily-scaled response as a sufficiently complex function of the constituent objects (where heavy least-squares computations become very probable).

(2) If we have enough diversity of weights, we may forget our quantitative response measures, except as indicators of greater or less, and go back to the first principles of Campbellian measurement. (After many rather tedious comparative weighings, we will reconstruct a conventional scale of weight, providing at the same time a calibration for our weighing instrument.)

(3) We may try various modifications of the mode of expressing our response, trying perhaps first the logarithm, the square root, and the square of the numbers provided by our colleague's scale, and then being guided in selecting new trial modes by such considerations as reduction of apparent interactions. (We are not likely to reach perfection in this way, but we are likely to greatly improve the behavior and understandability of our results. It might well be that they would become so clear as to suggest an approach to a physical theory of our measuring instrument.)

What are the pros and cons of these three ways out? To follow the first is clearly a counsel of desperation. To follow the second is a counsel of perfection. To follow the third is a counsel of empiricism and pragmatism. (A most significant aspect of the third way out is the great reduction in the labor associated with either of the other two if, as an initial step, the third is carried out with even partial success.)

In the weighing situation, we might hope to be wise enough to follow the counsel of perfection (preferably easing the rigors of the corresponding labor by taking a few steps down the third path first). But what of the opinion situation to which quadruplet weighing was an analogy? Various properties of the weighing situation are unlikely to

carry over. We are unlikely to be able to make highly precise comparisons, since sampling and questioning fluctuations will be relatively much larger than the instrumental fluctuations of a high grade weighing instrument. And we are not likely to have as free manipulation of the separate factors. (Formal education comes, for almost everyone, in standard doses, while brass weights could be made up in any desired size.) As a result of these and other considerations, the second way out is almost certain to be closed. We must then decide between the first and third ways. Since the third way, carried out at least part way, offers a labor-saving approach to the first way, there can be little doubt that it is almost always the best way to begin.

Thus the best beginning is almost certain to be the pragmatic, empirical one of trying different modes of expression in search of as much simplicity as we can readily obtain. Simplicity means: "If you change the quartz weight from Q_1 to Q_2 [if you change education from E_1 to E_2] the numerical change is as nearly as possible the same whatever brass weights, gold-plated weights, and aluminum weights are held constant [whatever reference groups, work groups and family groups are held constant]." When we deal with quadruplet weighing these numerical changes can be made extremely closely the same. When we deal with political-opinion formation, extreme closeness of agreement may not be attainable (though, for all we know today, it may be attainable). But if it isn't? There is still advantage in obtaining whatever simplicity we can by wise choice of mode of expression before allowing ourselves to be forced to deal with complex descriptions.

E4. SOME COMMENTS

For the case of two or more factors it is clear that one can describe in axiomatic form an approach to the ***joint measurement*** of the ***factors*** which would take over ***all*** the basic ideas and techniques of ***Campbell's treatment*** of the measurement of freely concatenable objects. And there would be no reason for giving such joint measurement lower social status or lower prestige than Campbell's fundamental measurement. The choice of a mode of expression which avoids all interaction, if this be possible, is the road to the best measurement from the high monastery view. How could it be more blessed?

One interesting change has taken place without explicit remark. We came in thinking we were to measure political opinion. We leave measuring the ***strength of the forces which mold*** political opinion (specifically those of education, of the reference group, of the work group, of the family group). Our measurements of political opinion are

cast in the mode which helps us in measuring the forces which mold it. We were led to this shift by the analogies with the situation in physical science as presented by Campbell, and driven to it by the logical exigencies of the situation. But it is not to the disadvantage of behavioral science that we have made this shift. More can, and will be, made from measurements of molding forces than from measurements of effects.

There is another point of interest and importance. Certain simple ways of changing the expression of a quantity should not be thought of as changing the mode. If expressing a response in square feet leads to an additive response, the same will be true when the response is expressed in square inches. Since one square foot is exactly 144 square inches, each expression in square inches will use a number exactly 144 times larger than that involved in the expression of the same quantity in square feet. If the one is additive, so is the other.

Similar remarks apply to expression in feet or in inches, and to latitude in degrees west of Greenwich or degrees west of Washington. Changes which involve adding the same constant to the numbers expressing all quantities, or multiplying all these numbers by some other constant, or both, change only the *expression*, not the *mode of expression*.

If we so desire, we can always readjust our expressions to have their zero at a convenient place by choosing an appropriate additive constant. Within one and the same mode of expression we can do this, and still be free to choose a multiplicative constant to meet one further requirement if we wish. Both kinds of freedom are convenient, and are frequently used.

E5. EXPRESSING COUNTED FRACTIONS

Counting sheep and goats, and reporting on the relative number of goats, still typifies much of behavioral science. And it is to be expected that this will continue to be so. Indeed, it should. As a consequence, the behavioral sciences have a very strong continuing interest in modes of expression of counted fractions, although they may appear to be unaware of this interest.

Are the conventional modes good ones? If not, in what direction should we go for better modes? For help in answering these questions we may look both to our own intuitions and to the philosophy built up in the previous sections. There is no doubt about which mode of expression is conventional, it is expression as a *percentage*, or, equivalently, as a decimal fraction. (On occasion, additional useful

information may be provided by giving the actual counts as common fractions, but these rarely enter further analysis as such.) What can we say against this mode?

First, our experience-molded intuitions tell us clearly that it is not a mode where equal numerical changes correspond to equally important changes. A change of 5% is not equally important across the scale. The difference, for almost all purposes except voting, between 1% and 6% is very much more important than the difference between 48% and 53%. Once we break down our idea that "percentages are the only proper mode," we come to feel quite clearly that we need to open out the scale for extreme percentages, as compared with percentages near 50%.

Second, we may draw a general inference from our discussion of the last few sections. If there are many possible factors to be changed, each with an effect which should be numerically nearly constant, we are likely to be in trouble if our scale has ends. For if we can move almost to the end of a scale, and still have a relevant factor which can change, one which should take us still further, then we may be stuck, may be unable to measure this factor as having its rightful effect, only and exactly because the scale ends. As a general consequence we should expect that scales which have a finite range are likely to give us trouble, unless all our observations tend to be safely away from any ends which are present. Hence the fact that percentages go only from one end (at 0%) to another (at 100%) suggests that, whenever even moderately extreme percentages are likely to occur, we are likely to have to "stretch the tails", while, if really extreme percentages occur, we may have to stretch hard enough so that there are no ends (at any finite values).

Third, experience with a rather wide variety of relative-number problems, varying from "how many were affected at this dose" to "how many of the pebbles are quartz", indicates that further analysis proceeds more smoothly and thoroughly when other modes of expression are used instead of "percentages".

Three modes with more extended tails are in more or less common use in various fields. While some have tried to provide deep philosophical reasons why one or another must be *the* correct one, all such discussion has proved ultimately unconvincing. These three modes are introduced here on the following reasons:

(1) in comparison with percentages, each stretches the tails (as compared to the middle);

(2) they differ enough among themselves that a choice among them is sometimes worth the effort;

(3) they have worked relatively well in many situations;

(4) they are commercially available (in comparable form) on convenient graph paper.

These three are not sacred, only useful. Their mathematical expressions are not simple in appearance, but tables and graph papers are freely available. (The proof of the pudding is in its eating, not its recipe.)

In order of successively longer tails, the names applied to the result of expressing relative numbers in these terms are:

(1) "anglits" ("sinits" or just "angles"), for which we see the following graph sheet labeled "Arc-sine transformation ruling" and numbered 32, 452, (31, 452 on thin paper), Figure 3.

(2) "normits" (or "probits") for which we see the then following graph sheet labelled "Normal ruling" and numbered 32, 451, (31, 451 on thin paper), Figure 4.

(3) "logits" for which we see the third following graph sheet labelled "Logistic ruling" and numbered 32, 450, (31, 450 on thin paper), Figure 5.

In my experience, much valuable insight into the behavior of bodies of data can be gained from the use of such sheets of graph paper, sometimes assisted by the use of tracing paper (or other means) to take off distances for replotting other graphs. Much of this insight can be gained, together with certain additions, by manipulating the corresponding numerical values.

To realize all these advantages we need not know the mathematical definitions of these modes of expression of relative numbers. We need only know how to read the scale on the graph paper, or how to enter and leave the tables, so that we may plot, or convert, our raw percentages. We are using these modes of expression as empirically useful tools, not as theoretically important constructs. If we can see that they serve our purposes, we shall certainly use them. If we can see that they do not serve our purposes, we shall use other modes without sadness or guilt. They are a simple tool.

E6. SOME NUMERICAL VALUES

We noticed earlier (in E4) that addition of a chosen constant, and multiplication by another would not take us out of a given mode of expression. The point of symmetry of a counted fraction is surely 50%,

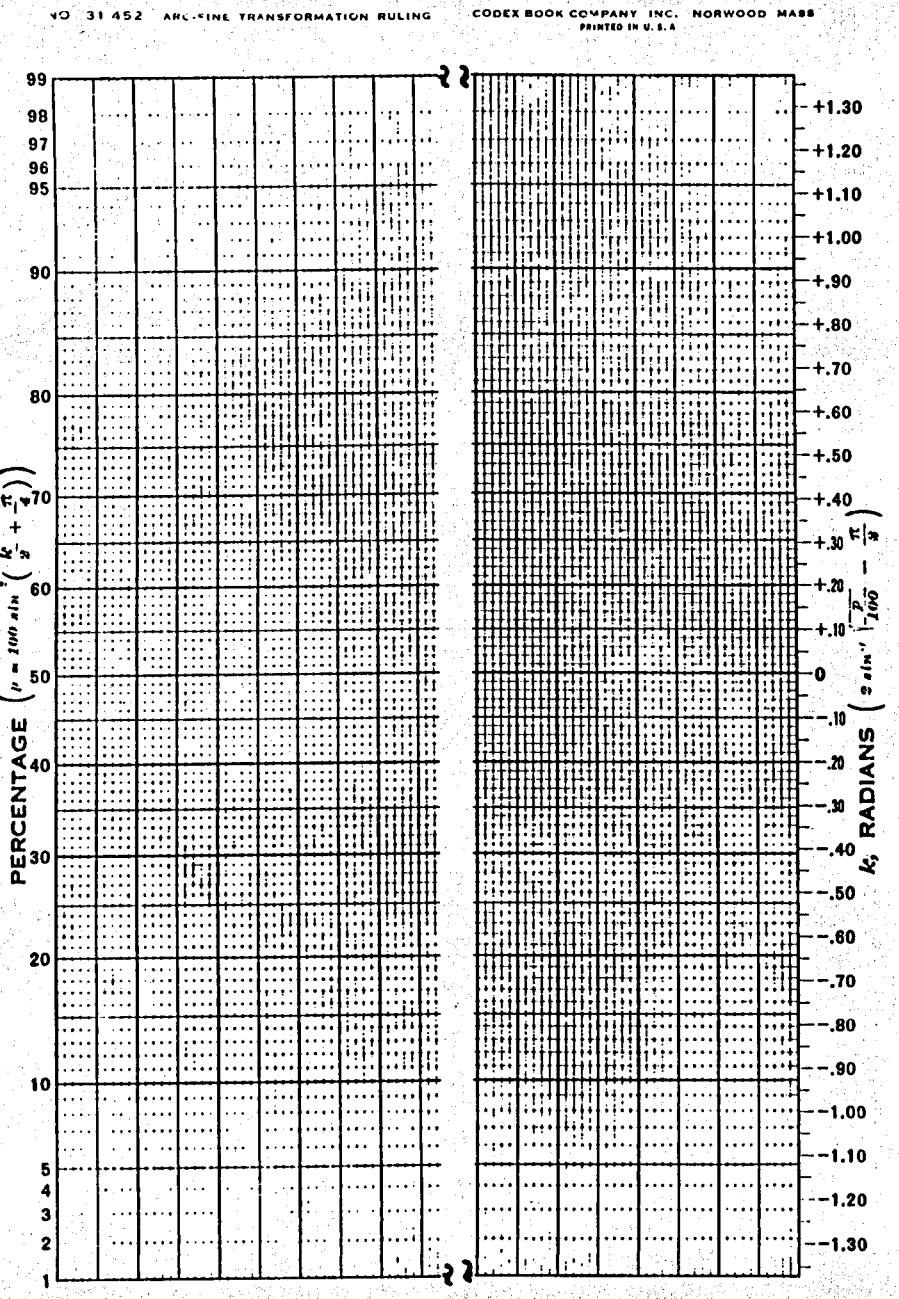


Figure 3. Arc-sine transformation ruling

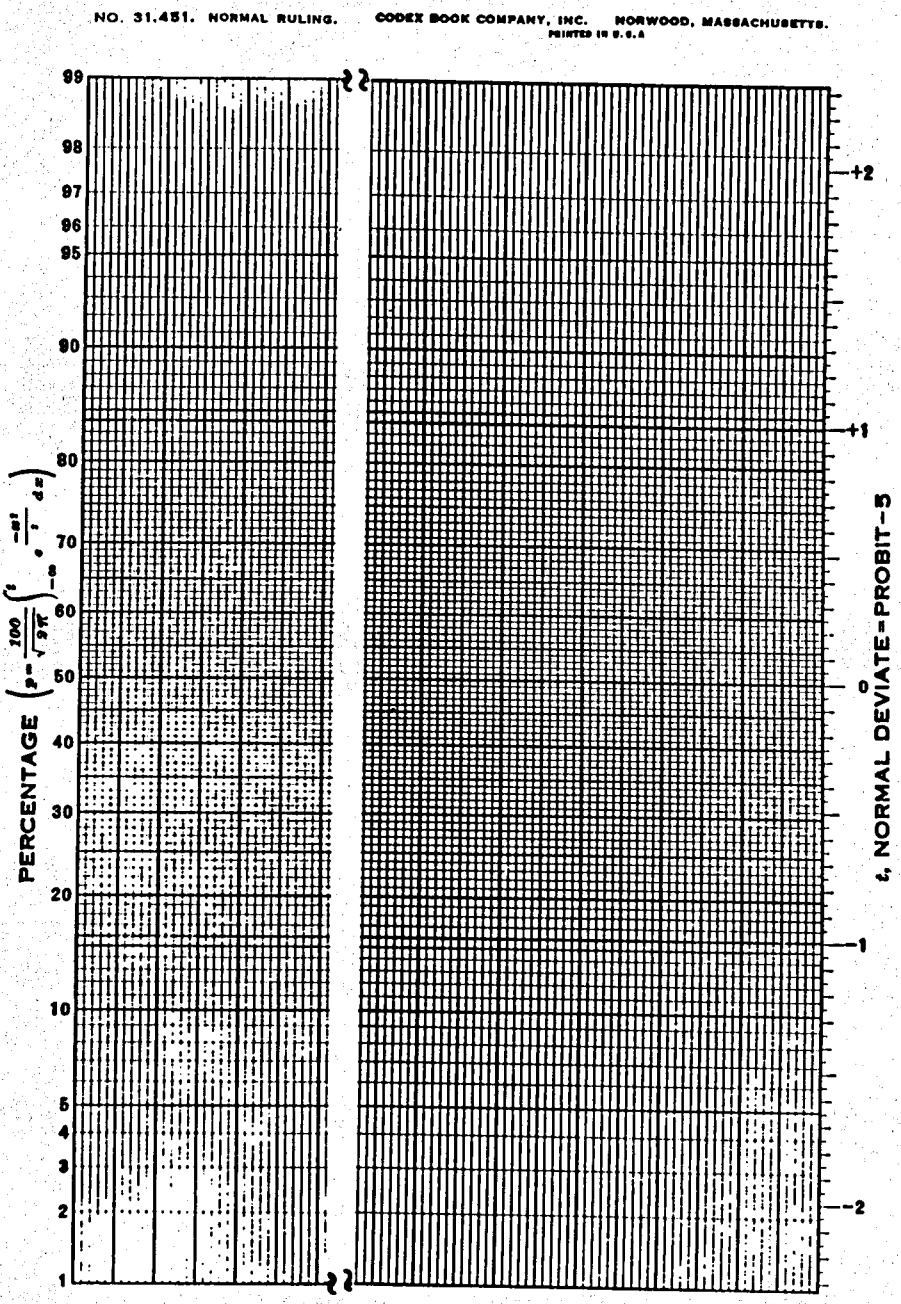


Figure 4. Normal ruling

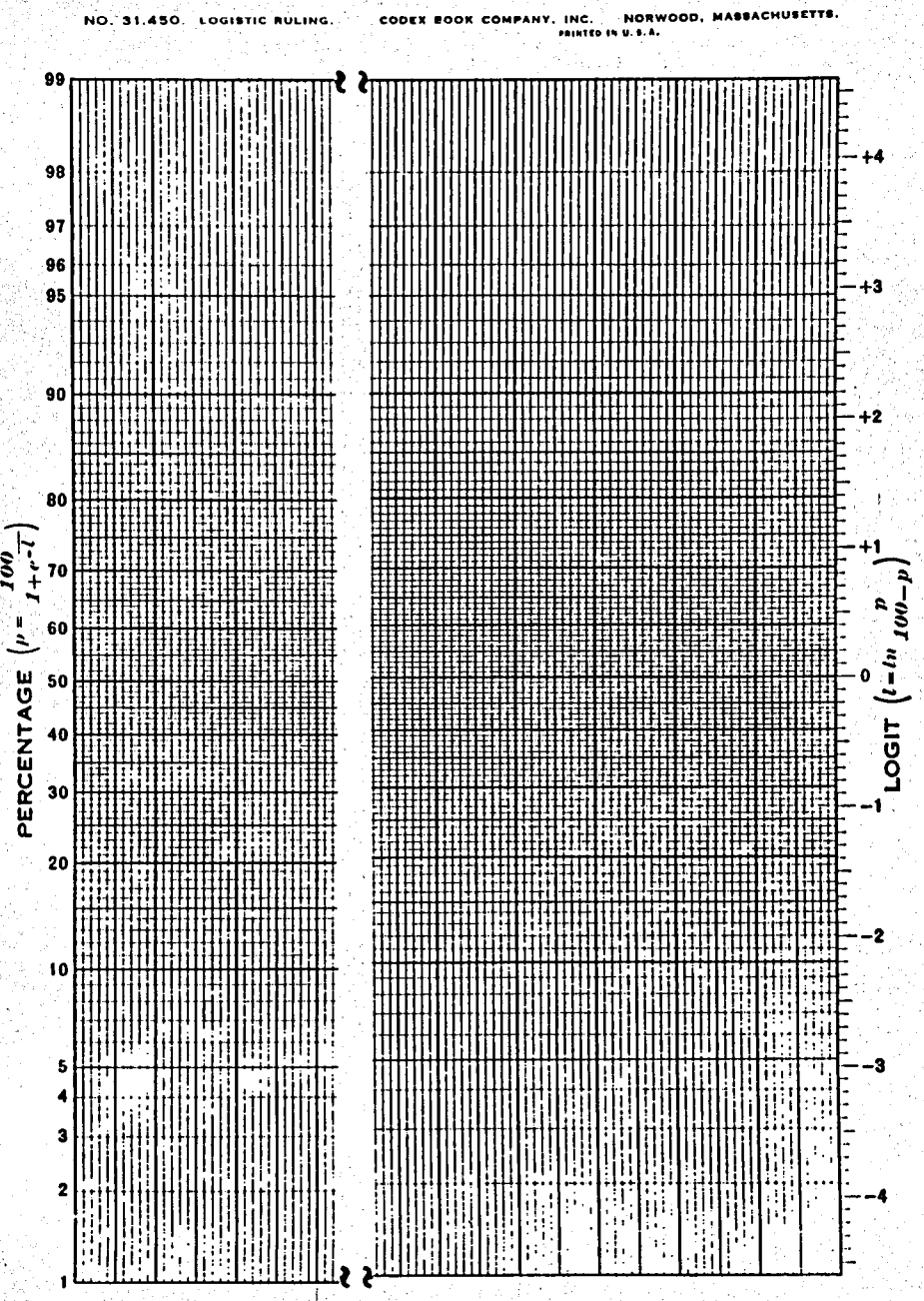


Figure 5. Logistic ruling

and it is thus natural to represent 50% by zero. Within the classical mode, the simplest such expression is

(1) $2 \cdot (\text{fraction}) - 1 = (\text{fraction}) - (\text{complementary fraction})$

which is −1 for 0%, 0 for 50%, and +1 for 100%. It will be instructive to choose additive and multiplicative constants for the other modes so that the resulting expressions match expression (1) in some sense. Matching at 0% or 100% is impossible, since two of the modes give $-\infty$ for 0% and $+\infty$ for 100%, so the best we can do is to match behavior near 50%.

Table 10 provides a brief table of values relating % to (1), which we will call "doubled fractions," and to center-matched expressions representing the three tail-stretching modes.

The values in Table 10 have been rounded to 2 decimals. Some will feel that this is ruthless. We shall see (in Section S3) that it is reasonable and gentle.

The effect of rounding must be judged by comparison with the fluctuation which was present before rounding. The least fluctuation that is commonly appropriate for a fraction is that for simple random binomial sampling. (Wisely stratified samples can, and indeed, on occasion, do have smaller fluctuations, but such situations are both infrequent and usually the result of careful planning. The results of most samples or "samples" show a *greater* variability than do the corresponding results for simple random samples. See, e.g., Kish 1957. Thus simple random variability is usually the least that needs to be feared.) For each of the modes of expression of fractions given in Table 10, we may take

$$\text{simple random sample variance} \doteq \frac{\text{numerator}}{\text{sample size}}$$

with the numerators given in Table 11.

The comparison of rounding variance with random sampling variance is made in S3 below. Insofar as anglits, normits, or logits are concerned, two decimals will surely suffice if the samples are not larger than, say, 1200, and will almost surely suffice for samples of sizes up to, say, 6000. Larger samples tend to have greater variability than that which corresponds to simple random sampling, so that two decimals will continue to serve in most cases. There may very occasionally be instances when more than two decimals will be appropriate. Table 36 in U4, below, gives values of anglits, normits, and logits to more decimal places, for use when necessary.

Table 10

Comparative values of various modes of expression for fractions for even %'s.

+	(1)	(2)	(3)	(4)	−	+	(1)	(2)	(3)	(4)	−
50%	.00	.00	.00	.00	50%	85%	.70	.78	.83	.86	15%
51	.02	.02	.02	.02	49	86	.72	.80	.86	.91	14
52	.04	.04	.04	.04	48	87	.74	.83	.90	.95	13
53	.06	.06	.06	.06	47	88	.76	.86	.94	1.00	12
54	.08	.08	.08	.08	46	89	.78	.89	.98	1.05	11
55%	.10	.10	.10	.10	45%	90%	.80	.93	1.03	1.10	10%
56	.12	.12	.12	.12	44	90.5	.81	.94	1.05	1.13	9.5
57	.14	.14	.14	.14	43	91.0	.82	.96	1.07	1.16	9.0
58	.16	.16	.16	.16	42	91.5	.83	.98	1.09	1.19	8.5
59	.18	.18	.18	.18	41	92.0	.84	1.00	1.12	1.22	8.0
60%	.20	.20	.20	.20	40%	92.5	.85	1.02	1.15	1.26	7.5
61	.22	.22	.22	.22	39	93.0	.86	1.04	1.18	1.29	7.0
62	.24	.24	.24	.24	38	93.5	.87	1.06	1.21	1.33	6.5
63	.26	.26	.26	.27	37	94.0	.88	1.08	1.24	1.37	6.0
64	.28	.28	.29	.29	36	94.5	.89	1.10	1.28	1.42	5.5
65%	.30	.30	.31	.31	35%	95%	.90	1.12	1.31	1.47	5%
66	.32	.33	.33	.33	34	95.5	.91	1.14	1.35	1.53	4.5
67	.34	.35	.35	.35	33	96.0	.92	1.17	1.40	1.59	4.0
68	.36	.37	.37	.38	32	96.5	.93	1.19	1.45	1.65	3.5
69	.38	.39	.40	.40	31	97.0	.94	1.22	1.50	1.74	3.0
70%	.40	.41	.42	.42	30%	97.2	.94	1.23	1.53	1.77	2.8
71	.42	.43	.44	.45	29	97.4	.95	1.25	1.55	1.81	2.6
72	.44	.46	.46	.47	28	97.6	.95	1.26	1.58	1.85	2.4
73	.46	.48	.49	.50	27	97.8	.96	1.27	1.61	1.90	2.2
74	.48	.50	.51	.52	26	98.0	.96	1.29	1.64	1.95	2%
75%	.50	.52	.54	.55	25%	98.2	.96	1.30	1.67	2.00	1.8
76	.52	.55	.56	.58	24	98.4	.97	1.32	1.71	2.06	1.6
77	.54	.57	.58	.60	23	98.6	.97	1.33	1.75	2.13	1.4
78	.56	.59	.61	.63	22	98.8	.98	1.35	1.80	2.21	1.2
79	.58	.62	.64	.66	21	99.0	.98	1.37	1.86	2.30	1%
80%	.60	.64	.67	.69	20%	99.2	.98	1.39	1.92	2.41	0.8
81	.62	.67	.70	.72	19	99.4	.99	1.41	2.00	2.55	0.6
82	.64	.69	.73	.76	18	99.6	.99	1.44	2.13	2.76	0.4
83	.66	.72	.76	.79	17	99.8	1.00	1.48	2.30	3.11	0.2
84	.68	.75	.79	.83	16	100%	1.00	1.57	∞	∞	0%

(1) doubled fraction = 2(fraction) − 1

(2) = anglit of fraction

(3) = modified normit = modified probit $= \left[\sqrt{\frac{2}{\pi}}\right]$ (normit of fraction)

$= \left[\sqrt{\frac{2}{\pi}}\right]$ (−5 + probit of fraction)

(4) = half-logit $= \left[\frac{1}{2}\right] \cdot$ (logit of fraction)

Table 11

Values of factor A in "variance $\sim A/n$," where n is the (simple random) sample size, for the modes of expression of Table 10

%	(1)	(2)	(3)	(4)	%
50%	1.00	1.0	1.0	1.00	50%
60	.96	1.0	1.0	1.0	40
70	.84	1.0	1.1	1.2	30
75	.75	1.0	1.2	1.3	25
80	.64	1.0	1.3	1.5	20
82%	.59	1.0	1.4	1.7	18%
84	.54	1.0	1.4	1.9	16
86	.48	1.0	1.5	2.1	14
88	.42	1.0	1.7	2.4	12
90	.36	1.0	1.9	2.8	10
91%	.33	1.0	2.0	3.1	9%
92	.29	1.0	2.1	3.4	8
93	.26	1.0	2.3	3.8	7
94	.23	1.0	2.5	4.4	6
95%	.19	1.0	2.8	5.3	5%
95.5	.17	1.0	3.0	5.8	4.5
96	.15	1.0	3.3	6.5	4.0
96.5	.14	1.0	3.6	7.4	3.5
97	.12	1.0	4.0	8.6	3.0
97.2	.11	1.0	4.2	9.2	2.8
97.4	.10	1.0	4.4	9.9	2.6
97.6	.094	1.0	4.6	11.	2.4
97.8	.086	1.0	5.0	12.	2.2
98%	.078	1.0	5.4	13.	2.0
98.2	.071	1.0	5.8	14.	1.8
98.4	.063	1.0	6.3	16.	1.6
98.6	.055	1.0	7.0	18.	1.4
98.8	.048	1.0	7.7	21.	1.2
99%	.040	1.0	8.9	25.	1.0
99.2	.032	1.0	11.	32.	0.8
99.4	.024	1.0	13.	42.	0.6
99.6	.016	1.0	18.	63.	0.4
99.8	.0080	1.0	32.	125.	0.2
99.9	.0040	1.0	56.	250.	0.1

Curiously enough it is only for column (1), "doubled fractions," an instance of the classical mode of expression for counted fractions, that two decimals may not entirely suffice for samples of less than 1000.

Greater precision for this expression is, however, only useful for fractions quite close to 0 or 1.

E7. AN EXAMPLE FROM CLINICAL PSYCHOLOGY

Our first example is drawn from Volume 4 of *Studies in Social Psychology in World War II* (Stouffer et al. 1950), where pages 512 to 538 present tables of frequencies of both individual answers and summarized scores for a variety of questions applied to 3,501 white enlisted men with no overseas service, and to 563 psychoneutrotic patients in Army hospitals. Table 12 presents the tables for four of the

Table 12

Comparison of score distributions for controls and psychoneurotics. (From Vol. 4 of *Studies in Social Psychology in World War II*, Stouffer et al. 1950, pp. 526-531.)

10. Sociability				12. Acceptance of solider role			
	Neurotic Patients	Cross Section	(Difference)		Neurotic Patients	Cross Section	(Difference)
			(original 2×4 tables)				
Summary Score	(%)	(%)	(%)	Summary Score	(%)	(%)	(%)
(3)	22	43%	(21)	(4,3)	16%	46%	(30)
(2)	33	41	(8)	(2)	25	23	(−2)
(1)	26	12	(−14)	(1)	32	20	(−12)
(0)	19	4	(−15)	(0)	27	11	(−16)
			(cumulative %'s)				
Break at	(%)	(%)	(%)	Break at	(%)	(%)	(%)
2.5	22	43	(21)	2.5	16	46	(30)
1.5	55	84	(29)	1.5	41	69	(28)
0.5	81	96	(15)	0.5	73	89	(16)
			(cumulative anglits)				
Break at	(<)	(<)	(<)	Break at	(<)	(<)	(<)
2.5	−.59	−.14	(.45)	2.5	−.75	−.08	(.67)
1.5	.10	.75	(.65)	1.5	−.18	.39	(.57)
0.5	.67	1.17	(.50)	0.5	.48	.89	(.41)
Mean			.53				.54

(%) = fraction, or differences of fractions, expressed in %, or difference in %.

(<) = fraction, or difference of fractions, expressed in anglits, or difference in anglits (cp. Table 10).

Table 12 (Cont'd)

14. Oversensitivity				3. Childhood fears			
	Neurotic Patients	Cross Section	(Difference)		Neurotic Patients	Cross Section	(Difference)
			(original 2×4 tables)				
Summary Score	(%)	(%)	(%)	Summary Score	(%)	(%)	(%)
(10,9,8)	29%	53%	(−24)	19	9	25	(16)
(7,6)	25	28	(3)	18-16	29	43	(14)
(5,4)	23	13	(10)	15-12	31	25	(−6)
(3,2,1,0)	23	6	(17)	11-0	31	7	(−24)
			(cumulative %'s)				
Break at	(%)	(%)	(%)	Break at	(%)	(%)	(%)
7.5	29	53	(24)	18.5	9	25	(16)
5.5	54	81	(33)	15.5	38	68	(30)
3.5	77	94	(17)	11.5	69	93	(4)
			(cumulative anglits)				
Break at	(<)	(<)	(<)	Break at	(<)	(<)	(<)
7.5	−.43	.06	(.49)	18.5	−.96	−.52	(.44)
5.5	.08	.67	(.59)	15.5	−.24	+.37	(.61)
3.5	.57	1.08	(.51)	11.5	+.39	1.04	(.65)
Mean			.53				.57

(%) = fraction, or difference of fractions, expressed in %, or difference in %.

(<) = fraction, or difference of fractions, expressed in anglits, or difference in anglits (cp. Table 10).

summarized scores and the results of converting their entries first into cumulative percentages and then into anglits. It is quite difficult, though perhaps possible, to examine the original tables carefully enough to detect the lawfulness and order that is actually present. It is certainly not possible to examine them closely enough to detect any possible deviations from the overall pattern.

The first step in getting a more quantitative hold upon the differences between cross-section and psychoneurotic patients is to focus attention not on the score values which fall in each of the 4 cells, but instead upon the three partitions or *breaks* which define these cells. The second part of the table, accordingly, lists cumulated %'s against breaks. (The %'s were cumulated downward from the top in this instance; cumulating up from the bottom would change only the signs of the anglits and their differences.) These cumulative %'s are then turned

into anglits with the aid of Table 9, with the result shown in the lower third of Table 12. The difference in anglits is everywhere nearly the same, both within and across tables, the greatest deviation from a differences of .53 being ±.13. This may be compared with a random sampling standard deviation of ±.04, which is obtained as the square root of the random sampling variance of

$$\frac{1}{3501} + \frac{1}{563} = .0020$$

Since the survey almost certainly involved some clustering, the observed deviations from scale to scale are plausibly consistent with no true differences.

The results for these four questionnaire scales may be quite completely summarized as being "an apparent shift of about 0.54 in anglit between the cross-section population and the psychoneurotic population." It is plausible to believe that this result is independent of the actual breaks used to form the given cells. (If the raw data were available, it would be easy to use all possible breaks, rather than only those given in *Studies in Social Psychology in World War II*, thus strenghtening the evidence on this point considerably. See U3 below for an example involving many breaks.)

It is probably worth remarking that the other items and scales also tended to show shift by an angle approximately constant for each item or scale, but differing from scale to scale. This is shown in Table 13.

This broader summary shows clearly the general extent of shift for each score, and also reveals some indication of whether the shift tends to vary systematically with the break chosen. In the writer's judgment, much insight into the data has been gained by conversion from cell percentages to cumulative anglits.

E8. AN EXAMPLE FROM ECONOMIC HISTORY

A second elementary example, of a quite different character, can be drawn from a discussion (between Landes and Gerschenkron) of the quality of industrialization in France and Germany during the early years of this century. The point at issue was the extent to which French industry was being carried on in smaller establishments. The evidence presented came from 1906-1907 census figures showing the number of establishments (and numbers of workers) in each of several size classes.

Table 13

Shifts in angle for two background items and 15 questionnaire scales. (Data from Stouffer et al. 1950, pp. 512-538.)

	Questionnaire item or scale	Differences	Mean Differences
	Age	.39, .28, .20	.29
	Schooling	.21, .51, .40	.37
	Rural-urban	.07, .02, −.06, −.05	.00
(2)	Relations with parents*	.16, .24, .21	.20
(3)	Fears*	.44, .61, .65	.57
(4)	Neurotic symptoms*	.59, .70, .60	.63
(5)	School adjustment*	.26, .10, .19	.18
(6)	Fighting behavior*	.29, .31, .30	.30
(7)	Participation in sports*	.22, .48, .54	.41
(8)	Emancipation from parents	.03, .12, .21	.12
(9)	Mobility	.03, .12, .19	.11
(10)	Sociability	.45, .65, .50	.53
(11)	Identification**	.12, .16, .19	.16
(12)	Acceptance***	.67, .57, .41	.54
(13)	Worrying	.52, .50, .41	.48
(14)	Oversensitivity	.49, .59, .51	.53
(15)	Personal adjustment	.65, .76, .74	.72
(16)	Psychosomatic complaints	(See Appendix)	(larger)

* as a child.

** with war effort.

*** of soldier role.

Table 14 illustrates the conversion of the size-class data into percentages below and above certain size breaks, whose choice is determined by the way in which the original tabulations (Landes 1954) were made. Table 15 compares the difference between the size distributions in France and Germany when expressed (i) as a difference in percentage, or (ii) as a difference in logits. Over the range of size breaks at hand, the differences in percentage vary from as much as 2 or 3% to as little as 0.02%. The differences in logits are, by contrast, about the same at each of the break points.

For either "industry and mining" or "commerce" the distributions of establishment size in Germany is shifted about 1.1 logit toward bigger establishments. This is a simple statement which sums up the bulk of the figures presented. There seems to be no description of the

Table 14

Percentages of establishments of given sizes in France (1906) and Germany (1907) expressed in percentages in different ways. (Date from Landes 1954)

Size Class	Size Break	% in class France	% in class Germany	Relation to break (smaller and larger in %) France	Relation to break (smaller and larger in %) Germany
----------Industry and mining----------					
1-10		98.02%	94.60%		
	10.5			98.02 and 1.98	94.60 and 5.40
11-50		1.60%	4.06%		
	50.5			99.62 and 0.38	98.66 and 1.34
51-200		0.30%	1.07%		
	200.5			99.92 and 0.08	99.73 and 0.27
201-1000		0.08%	0.24		
	1000.5			99.99 and 0.01	99.97 and 0.03
1000 up		0.01%	0.03%		
----------Commerce----------					
1-10		98.95%	97.01%		
	10.5			98.95 and 1.05	97.01 and 2.99
11-50		0.97%	2.74%		
	50.5			99.92 and 0.08	99.75 and 0.25
51-200		0.07%	0.22%		
	200.5			99.99 and 0.01	99.97 and 0.03
201 up		0.01	0.03%		

(Rounding of some values adjusted for consistency.)

Franco-German relationship in terms of the percentages shown in the left-hand side of Table 15 which is even remotely simple.

If similar behavior (i.e., roughly constant displacement on the logit scale) were shown by intercensal comparisons within these countries, and between other pairs of countries, this mode of expression might prove quite useful in compressing extensive tabulations to much more easily perceivable figures.

Landes also gives figures for a number of individual industries. Figure 6 shows the relation of logit to breaking point for size of establishment for six of these. The three on the left, inorganic chemicals, electrical machinery, and water transport, behave quite similarly. Not only spacings but slopes are about the same. Although the three on the right, mining, chemicals (as a whole), and textiles, behave quite differently from one another, their behavior in either country alone is quite simple to describe, as is their comparative behavior in the two countries.

Table 15

Relative numbers of establishments below and above certain sizes for France (1906) and Germany (1907) expressed in percentages and logits. (Data from Landes 1954)

Size break (personnel)	Percentages (smaller and larger) France	Germany	diff.	Logits* France	Germany	diff.
	----------Industry and mining----------					
10.5	98.02 and 1.98	94.60 and 5.40	3.42%	3.90	2.86	1.0
50.5	99.62 and 0.38	98.66 and 1.34	0.94	5.57	4.30	1.3
200.5	99.92 and 0.08	99.73 and 0.27	0.19%	7.08	5.91	1.2
1000.5	99.99 and 0.01	99.97 and 0.03	0.02%	9.42	8.22	1.2
	----------Commerce----------					
10.5	98.95 and 1.05	97.01 and 2.99	1.94%	4.55	3.48	1.1
50.5	99.92 and 0.08	99.75 and 0.25	1.17%	7.12	6.00	1.1
200.5	99.99 and 0.01	99.97 and 0.03	0.02%	9.17	8.22	1.0

(*calculated using more significant figures than given in percentages.)

(Some might try to argue at this point that whatever simplicity we have gained by using logits might equally well have been gained by merely plotting cumulative probabilities on a logarithmic scale. The instance of inorganic chemicals in Germany shows that this is not the case. Here 58% of the establishments fall above the first size break (have 11 or more personnel), and the use of the difference in logarithms of the two percentages is about 0.9 units different from using the logarithm of 58% alone. The symmetry of the logit, as between p and $1-p$, is here of great value in producing results which are simple and easily describable.)

This example is also just a hint, but a rather strong one. What is it that this example exhibits? It is something more than a mere production of a summary figure, such as the difference between the average number of personnel in establishments in the two countries would have provided. If we knew such a difference in averages, we could not predict any detail of one size distribution, given the whole of the other. But if we were sure that, throughout the size range, the differences in logits were 1.1 logit units, we could make individual predictions. Thus, at another place, Landes gives the percentages of establishments with 5 or less persons in the two countries. In commerce

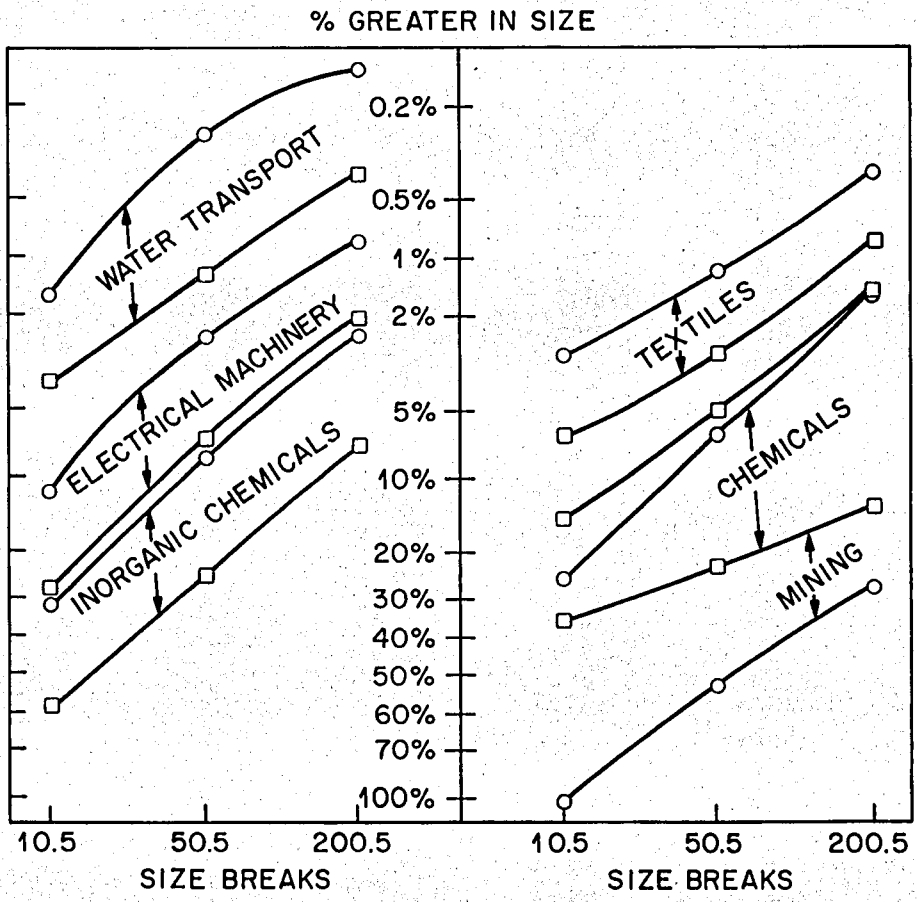


Figure 6. Size distribution of certain sorts of French and German enterprises (logit scale).

and finance this is 92.59% in Germany, corresponding to a logit of 2.53. Adding 1.1 to this yields a logit of 3.63, corresponding to 97.42%. Landes's figure is 97.06%, to which 97.42% is a moderately good extrapolation. (The actual difference is 0.97 logit unit, say 1.0 in comparison to 1.1.)

One of the main morals of this example is the great advantage of the choice of a mode of expression which permits a single number to transmit information about many details. This can *only* be done, in effect, by choosing a mode of expression for the details (here the comparisons at various size breaks) in such a way they all have the same numerical value.

E9. THE "PERCENTAGE FALLACY"

The principle that splits should be expressed in terms of percentages, fractions, or some equivalent, even though it be accepted tacitly rather than explicitly, can lead to conclusions which are poor science, though they may perhaps be good engineering. It is important to understand this class of situation, since the engineering conclusions may be quite different from the scientific ones.

Let us begin with a hypothetical example involving houseflies and two insecticides which can be used alone or in a mixture, and whose lethal effects arise by entirely different routes. More specifically, if a fraction p_i of the flies in question will survive a dose a_i of the first insecticide, while a fraction P_j will survive a dose B_j of the second insecticide, the fraction of flies surviving the mixed dose $a_i + B_j$ shall be $p_i P_j$; survival shall be independently and at random. From a scientific point of view, the second insecticide is equally effective in the presence or absence of the first; a given dose kills a given fraction. How could constant effectiveness be more clearly expressed?

Let us take a numerical example, and put it in mournful percentages. Consider 4 doses of the first insecticide with lethalities as follows:

Dose	Lethality	Survival
a_0	0%	100%
a_1	50%	50%
a_2	90%	10%
a_3	99%	1%

Consider only one dose, B, of the second insecticide, with 75% lethality. Then, cutting each % survival to a quarter

Dose	Survival	Change in % survival due to dose B
$a_0 + B$	25%	75%
$a_1 + B$	12.5%	37.5%
$a_2 + B$	2.5%	7.5%
$a_3 + B$	0.25%	0.75%

Expressed in difference of %, the effect of dose B of the second insecticide falls off as the dose of the first insecticide increases. How should we interpret this result?

At one extreme we might desire to control the insect population of an inhabited area by applying insecticide to the more important sources. Knowing that other sources will surely contribute insects, we should be concerned with the reduction in total numbers. While % reductions are not necessarily directly applicable, more detailed computation being needed, they may reflect the engineering usefulness of dose B of the second insecticide in the presence of varying doses of the first. The change in % *may* reflect an engineering truth, although it clearly distorts scientific truth atrociously and dangerously in such an example.

Let us ask what sort of action by a second insecticide would give a constant increase in % reduction, and see if such a "constancy" is anything against which we could care to compare possible or actual behavior. At dose a_3, where 99% are already dead, such a "constant % change" agent could at best have 1% effect. One example would thus be a dose of an agent which killed *only 1%* of the flies that would otherwise survive when the first agent was at the ineffective dose a_0, but would kill *all* the flies that would survive when the first agent was at the very effective dose a_3.

To make the example more behavioral, let us try to kill illiteracy, rather than insects. Would we really feel that an agent (probably some sort of intensive literacy program) which would (a) produce 1% literacy in an otherwise wholly illiterate culture, (b) increase literacy by 1% in a 50% literate culture, or (c) eliminate every single case of illiteracy in a culture already 99% literate, represented a force of equal strength in all three cases? Surely either (a) or (b) is much easier to accomplish than (c). Surely the constant % change standard is not a satisfactory measure of impact (even though it may sometimes be a satisfactory measure of result).

Now one might say that well-trained behavorial scientists would not be beguiled by this fallacy. Let us look at Herbert Hyman's book on survey design and analysis (Hyman 1955) in which very many issues have been carefully thought through and worked out. Turning to page 297, we find Hyman discussing the relationship of sex and campaign interest to the probability of voting in the next election in these words: "That is, the influence of the varying social roles— and social responsibilities— of men and women was most pronounced among those who expressed least interest in the coming election." What do the facts really say, and how can we let them speak out?

We can do quite well by letting the data speak in logits and

Table 16

Voting — not voting by sex and expressed interest in the election. (Hyman 1955, page 297)

	men			women			Difference	
Expressed interest	(No.)	% voting	half-logit voting	(No.)	% voting	half-logit voting	%	half-logit
Great	(449)	99%	2.30	(328)	98%	1.95	1%	0.35*
Moderate	(789)	98%	1.95	(852)	87%	0.95	11%	1.00
None	(56)	83%	0.79	(238)	44%	−0.12	39%	0.91

*This value subject to extra uncertainty.

(i) because of higher standard error due to simple random sampling (Table 11 leads to ± .31 instead of ± .14 and ± .19, respectively) and

(ii) because of large effects due to rounding observations to integer %. (If %s were 99.4 and 97.6, difference in half-logits would be 0.70 instead of 0.35.)

differences of logits. Table 16 sets forth the data, in % and in half-logits (from column (4) of Table 10). As the footnote makes abundantly clear, the data reproduced by Hyman is entirely consistent with a constant shift of about 0.9 half-logit.

Thus there is no basis for assuming that sex role and responsibility differences have had a different impact on voting behavior at various levels of interest. The sociologically meaningful conclusions, then, are:

(1) sex, perhaps acting through role and responsibility differences, corresponds to a difference in probability of voting of about 0.9 half-logit in the situation studied;

(2) this shift *may* not depend upon level of expressed interest in the election at all, though possible changes in shift have not been measured with great precision.

The conclusions reached by Hyman are appropriate to the political engineering (= the practicing politician).

(The reader may be interested in applying similar analyses, using logits, to others of Hyman's examples, including Table 18 on page 291, Table 20 on page 293, Table 25 on page 296, noting standard errors of differences when appropriate.)

For further discussion of this general subject, see Appendix U.

*E*1. THE EXAMPLE FROM ECONOMIC HISTORY*

In E8 we compared the size distribution of establishments in varying sorts of business and industry in France (1906) and Germany (1907). When expressed graphically, the results for (i) industry and mining, and (ii) commerce, the results appear as in Figures 7 and 8, if we confine our attention in each case to the top 3% (in size) of the establishments. Essentially nothing is to be learned from the bar diagram. The logit equal-area diagram shows something, but not much.

*E*2. THE VOTING EXAMPLE*

In E9 we took up, as an example of "the percentage fallacy", data on the relation of sex and reported interest in an election to reported intention to vote. The results may be displayed is in Figure 9 (in percent) or as in Figure 10 (in logits). In this example the change in impact is striking.

F. PROCEDURES OF COMBINATION

"In union there is strength" is a motto sometimes neglected and sometimes misinterpreted so far as the analysis of counted data is concerned. In those branches of modern statistics which deal with measured data, especially measured data which comes from experiments, much use is made of strength through union, though this fact is kept relatively secret. The purpose of this chapter is to explain some of the principles and, to a lesser extent, illustrate their potential uses in connection with counted data.

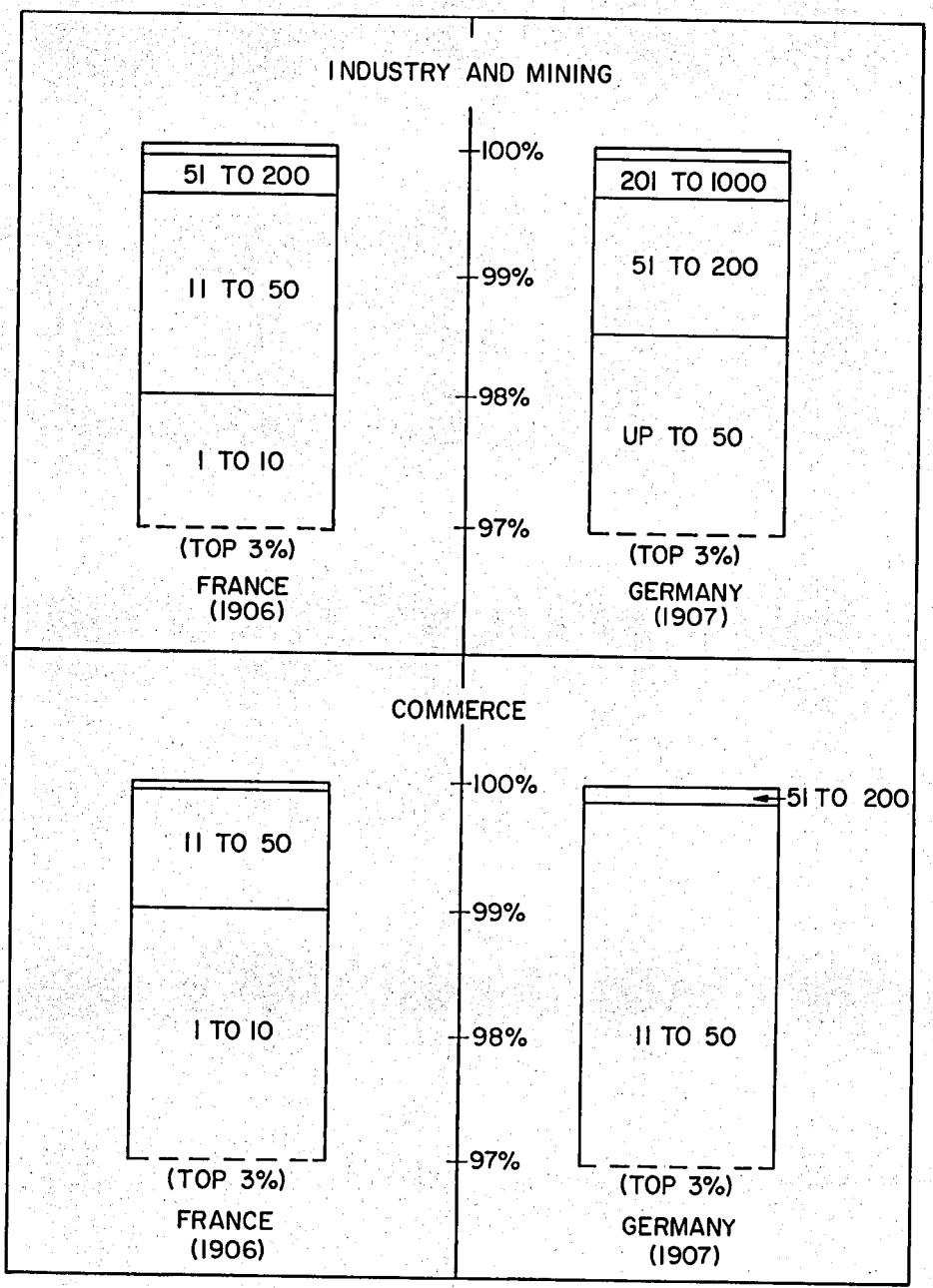


Figure 7. Size distribution of establishments of various sizes expressed in percent. [Data from Landis 1954]

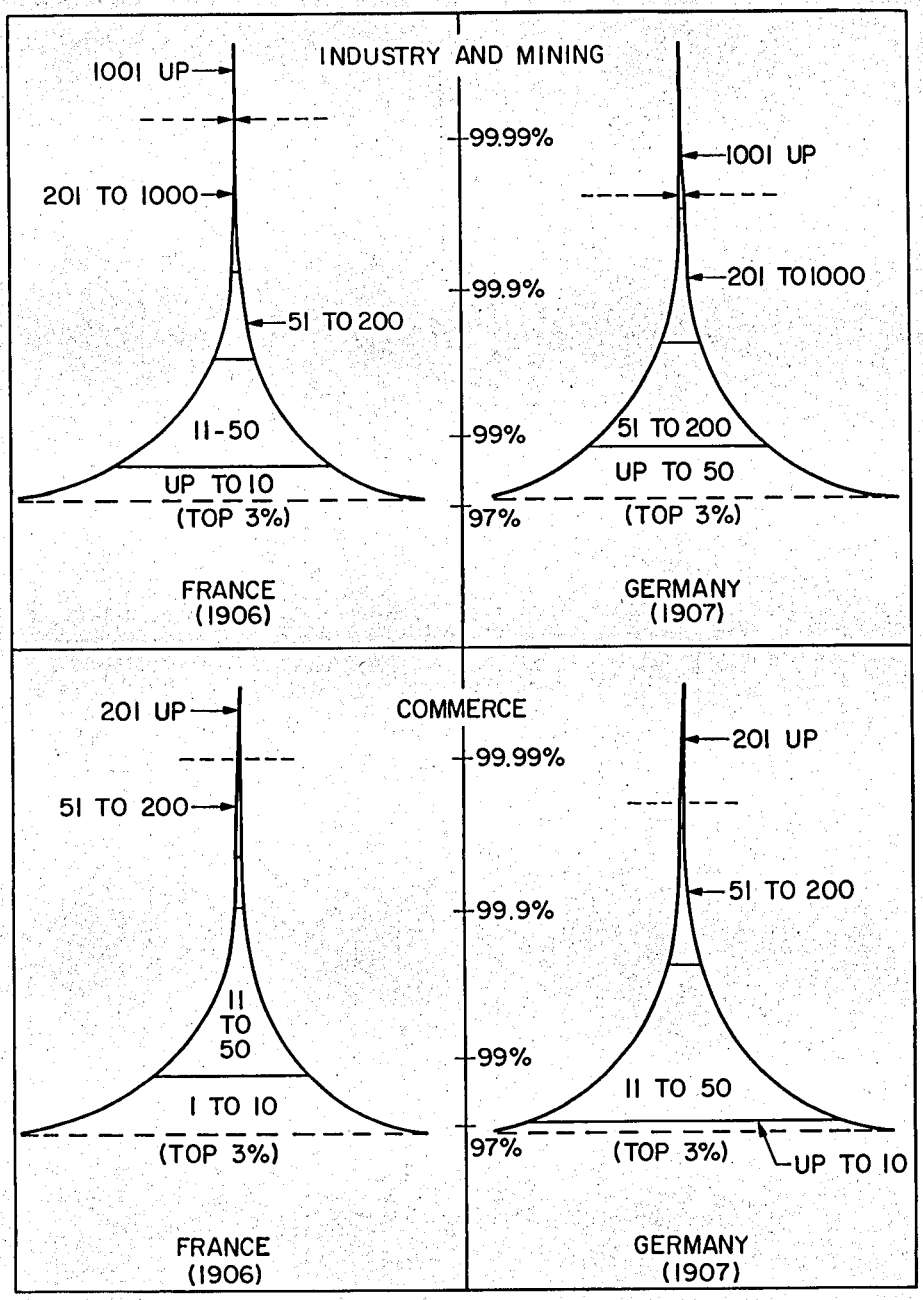


Figure 8. Relative number of establishments of various sizes expressed in logits. [Data from Landis 1954].

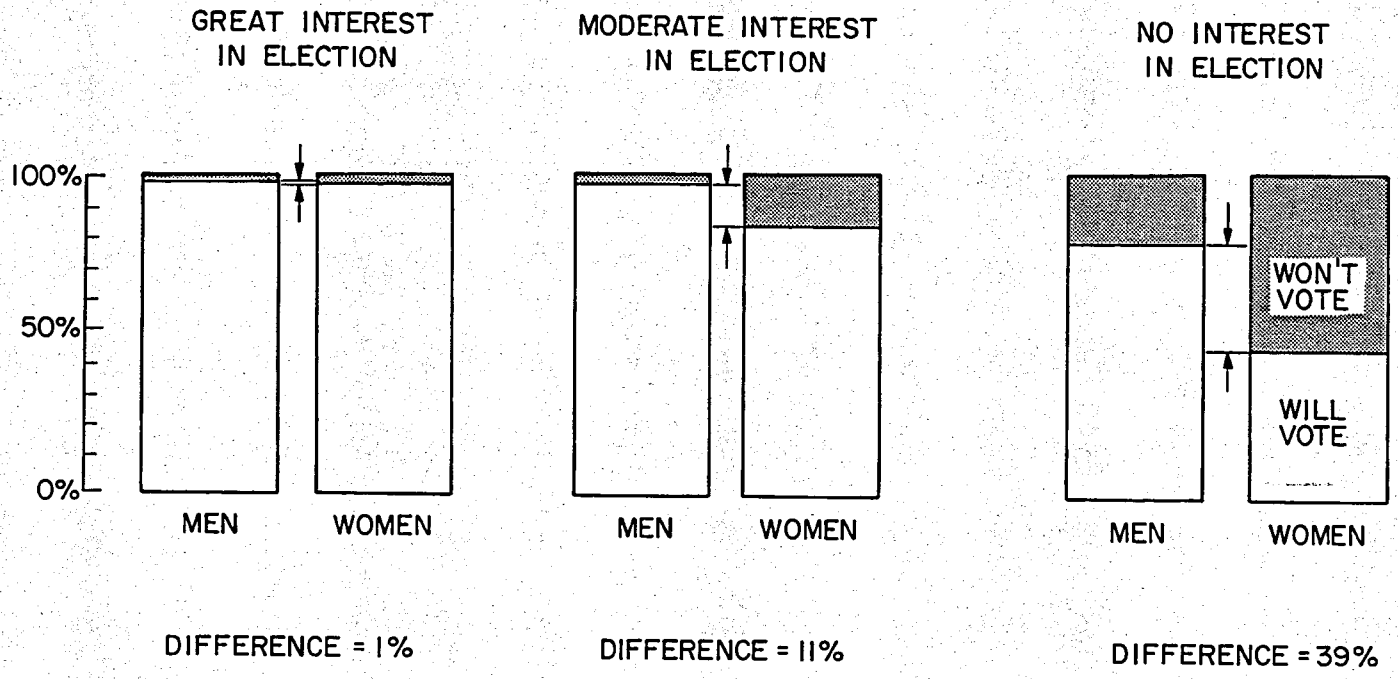


Figure 9. Relation between sex, expressed interest in campaign, and expressed intention to vote [Lazarfeld Berelson, Gaudet 1948, Hyman 1955] displayed in *percent* to emphasize relation of sex differences in expressed intention to expressed interest.

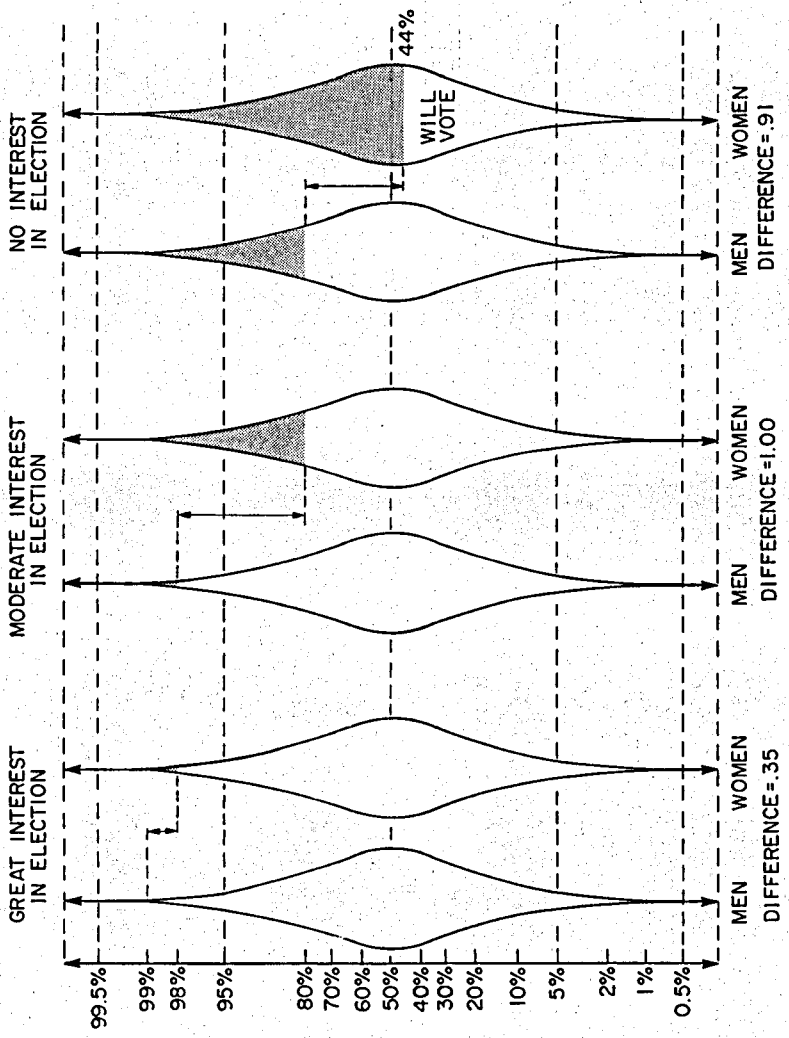


Figure 10. Data of Figure 9 expressed on *half-logits* to emphasize relation of sex difference in expressed intention to expressed interest.

F1. BORROWING STRENGTH

One basic notion is that of ***borrowing strength***. From a highly puristic standpoint such actions appear logically unsound, but from the respective standpoints of practical needs, and collated human experience, such actions are necessary and wise. But what action? Let us illustrate the situation thus. Suppose some phenomenon of interest in some behavioral science context has appeared in some British observations, observations which are of high quality but not extensive enough for us to have great confidence in the appearances. What is the natural thing to do? Certainly to look for other material. When found, it may be German, Italian, Spanish, or Texan. Do we not use it because it is not British? We shall surely make an attempt to use it, unless we have strong reasons to expect substantial and meaningful differences in behavior in one of these other societies. We will try to "borrow strength" from other situations (so long as it is not unreasonable that they be similar) ***even*** if we are only concerned with Britain. (If we were concerned with general principles alone, we should not be "borrowing". Instead we should be "broadening the basis of our inference".)

There are many ways to express the "philosophy" just illustrated. One may say, if he wishes: "When unable to measure individual situations precisely enough, guide yourself (in whatever individual situation you may be) by the more precise measure available for the average situation." When the individual situations are people, this statement describes the activity of the life insurance industry. In that instance, following such a "philosophy" is generally recognized as wise. But when applied to combining German and Spanish data with British data in order to draw conclusions about Britain, it is not quite so respectable. It is probably right that it should not be ***quite*** as respectable, but it would be a shame if a slight loss of respectability entirely prevented such combination.

Another way to describe this "philosophy", using statistician's jargon, is to say: "Unless the 'interactions' are substantial, depend on the 'main effects'." Here the main ***effects*** are average behaviors over various instances or situations, or estimates of such average behaviors. Much of the functioning of the analysis of variance revolves around the concept of main effect, which, like many useful concepts, is precise enough (as the arithmetic average of what goes on in various situations) not to disturb theorists (or mathematicians), and still flexible enough to serve usefully when only limited data is at hand.

It is my impression that those for whom cross-tabulation is the only analysis do not borrow strength nearly as much as they might. It is all too easy, once the cross-tab is before one, to try to put into words

only the *differences* between the phenomenon exhibited in the various columns, to omit explicit recognition of the ways in which the columns exhibit similarities, in which the columns reinforce one another. One cause of such omissions may be the misconception that the proper way to let the columns reinforce one another is to add up across columns, and look at the cross-tabulation with one less breakdown. (This may sometimes be proper, but it is often quite improper.)

To "borrow strength" it is often necessary to have the plausible effects of sampling fluctuations quite firmly in mind, to think of each piece of information as fuzzy. This can be uncomfortable to some. How great a role this consideration plays in the underuse of strength borrowing is also hard to judge.

F2. *"POOLING WITHIN"*

The general discussion of the last section would certainly do little more good than the average Sunday sermon if there were nothing to say but such great generalities. Fortunately, this is not the case. There are simple technical devices which make use of the broad principle and let us do things we could not otherwise do. One of these is described by the words "pooling within". The basic idea is to gather quantitative indications from "within" various parts of the data and then "pool" these indications into a single overall indication (which we may then sometimes be forced to accept as the most reasonable indication of what is going on "within" each portion).

A convenient and illuminating counted-data example arises when we have counts of a's and A's, separated as to b's and B's within each of a number of portions of the data, which we shall designate as $C_1, C_2, \ldots, C_6$. Table 17 shows some hypothetical numbers. No one of the six two-by-two tables really gives strong evidence for more a's among b's than among B's. (In fact, continuity-corrected chi-squares are all trivially small.) Yet each offers some evidence and, if we may combine all these bits and pieces together, we will have useful evidence.

In order to make a quantitative combination, we must measure, in some way, the shift in fraction of a's (from b's to B's) *within each* of the six portions. Two of many possible modes of expressing this shift are put to use in Table 14. First, we may consider merely

$$(\%\ a\text{'s among } b\text{'s}) - (\%\ a\text{'s among } B\text{'s})$$

which varies between +4.7% and +13.7% with a mean of +8.9%. (If we use Student's t to set 95% confidence limits, we find that the mean difference in % lies between 5.5% and 12.3% with 95% confidence.)

This second comparison is made on the basis of the difference in anglits, whose average value is +.24 and for which a 95% confidence interval extends from +.19 to +.29. (We may regard this latter interval as the narrower one. In realistic examples the writer would expect the use of anglits, or of normits or logits, to provide, by and large, somewhat more searching analyses than the use of %. The numbers treated here are purely hypothetical, and thus provide no evidence of this. For the point presently being made, the particular mode of expression used for the indications provided by each of the various two-by-two tables in *this* example is not important.)

Table 17

Hypothetical example of "pooling within." The influence of "*b* or *B*" on the relative number of *a*'s within, each of six sections of data $C_1, C_2, \ldots, C_6$

(1)	(2)	C_1	C_2	C_3	C_4	C_5	C_6	(within)	total
b	*a*'s	10	15	22	37	15	18	—	117
	A's	82	61	40	10	2	1	—	196
B	*a*'s	1	2	7	91	63	97	—	261
	A's	15	19	25	40	17	12	—	126
b	%*a*	10.9	19.8	35.5	78.7	88.2	97.7	—	37.4
B	%*a*	6.2	9.5	21.8	69.5	78.8	89.0	—	67.4
diff. in	%*a*	4.7	10.3	13.7	9.2	9.4	5.7	(+8.9)	−30.0
b	anglit	−.90	−.65	−.29	.61	.87	1.11		
B	anglit	−1.07	−.94	−.58	.40	.61	.90		
diff. in	anglit	.17	.29	.29	.21	.26	.21	(+.24)	−.71

Thus, "pooling within" would, in this hypothetical instance, bring out clearly the positive relationship between *b*-ness and *a*-ness which no one of the individual tables could demonstrate. And what would have happened if the oversimple approach of combining all six into a "total" two-by-two table had been used? Shockingly strong evidence of a relation between *a*-ness and *b*-ness. Shockingly strong because of the fact that such carelessly pooled evidence points in the wrong direction, showing relatively more *a*'s among the *B*'s than among the *b*'s. In this instance, which was no doubt carefully loaded, the "pool within" and "look at the combined table" approaches have come to quite different answers. We cannot try to blame this on sampling fluctuations. For the results would still have been in the same opposite directions if all the

original numbers were exactly 100 (or exactly 1000) times as large. We have no recourse but to blame this difference on the difference between the questions to which the two modes of analysis were responding.

We may phrase these two questions in a variety of ways, but the following will serve us well enough:

(1) "How are a-ness and b-ness related within a typical subdivision C_i of the data?"

(2) "How are a-ness and b-ness related without regard to possible subdivision by C_i?"

There is a great difference between such questions. This is specially true when the subdivision is by such variables as sex, age, geographical location, occupation, or socio-economic status. If we ask one question by mistake for the other, we may make a most serious error. It is important to us not to be forced into the position of having ***faute de mieux*** to answer the wrong one. As Lazarsfeld (1958, page 121) points out, it has been found that the more fire engines that come to a fire, the more damage. Yet unable to escape answering the wrong question is the position many have felt themselves in. If a subdivision makes the number "too small" for individual analyses, and "pooling within" is an unknown technique, then there is little escape · · · the second question is the one that will be answered, whether or not it is the correct one.

The use of "pooling within" may be either essentially qualitative or almost completely quantitative. It can be used, not only to answer the correct question, but to answer this question in a more searching and revealing way, especially in more complex situations. To do the latter thoroughly and well, it must be thoroughly quantitative, and as the examples of Sections E7 and E8 illustrate, it must often be aided by a wise choice of the mode of expression.

(It may be helpful to notice in passing that the basic concept of "borrowing strength" is not restricted to the assessment of directions and amounts of individual differences. Finding and using a ***pooled*** estimate of error, so characteristic of the analysis of variance, is an instance of "borrowing strength" to estimate how large typical errors and fluctuations are likely to be.)

F3. "ADJUSTED FOR"

In the physical sciences one frequently comes across the words "corrected for". They mean that the effects of some variable irrelevant to the topic of immediate discussion have been removed as well as we know how, and thus been kept from interfering, either systematically or

irregularly, with our study of the immediate topic. Usually the correction is based on theory, though there are circumstances where it is based on experience. Physical scientists place results "corrected for so-and-so" in the *highest* of the social classes into which results may fall.

The social class next below this contains, in their view, results "adjusted for so-and-so". The words "adjusted for" imply an empirical (and therefore undoubtedly somewhat incomplete) compensation for the effects of some variable irrelevant to the immediate topic. The basis for selecting the form of compensation used will have been empirical, quite often consisting of the single body of data at hand. In a very real sense, this too is usually an example of "borrowing strength", one which operates in a more subtle way. Thus, for example, in dealing with measurements, we may borrow information about the nature and extent of the apparent effect of the irrelevant variable from "interactions" and use this information to adjust "main effects". (This procedure is called the analysis of covariance.) In a sense, too, such methods of adjustment, in which constants are fitted to the data, could be considered as examples of data-guided analyses. No statistician thinks of them this way, however, because formal methods of "adjustment" were developed and became standard long ago.

Why should the behavioral scientist be concerned with methods of "adjusting for", when he may use cross-tabulation, especially cross-tabulation strengthened by "pooling within"? To this question there are various answers, some involving the relative efficacy of cross-tabulation and adjustment, while others relate to the very practical fact that adjustment for many irrelevant variables is much more feasible than is equally many-way cross-tabulation, with or without "pooling within". Thus simultaneous adjustment for nine variables is a substantial computing job, but in no wise out of hand with modern equipment, while an adequate nine-way cross-tabulation is almost (though perhaps not quite) unmanageable.

Let us return to the example of the Atlantean-Americans vs. the Muan-Americans, and their income distributions. Let us suppose that we can measure the social or occupational "class" of the individuals studied on a very much more finely divided scale than either of those used in Table 7. If we were to repeat this table, using the narrowest cells available, we would undoubtedly face very small numbers in each cell, with correspondingly large fluctuations in average incomes. A direct cross-tabulation approach with very fine cells would not give final answers. But it would be a first step (though not necessarily the simplest one) toward good answers. For when these narrow-cell average incomes were plotted, separately for Atlanteans and Muans, against our measure of social status, some regularity could be seen to

underlie the fluctuation. And after reasonable rescaling of the social status scale, if necessary, this relationship would, for the situation contemplated, be roughly linear for each group of immigrants, the two linear relations having approximately the same slope.

Having now a convenient measure of social status, and an approximately common slope (regression coefficient) of average income on social status, we can proceed to adjust each individual income for the (approximate) *linear* effect of social status by forming

$$\text{(actual income)} - \text{(slope)} \times [\text{(actual status)} - \text{(reference status)}]$$

for each individual. Doing this:

(1) will not distort or make unfair the comparison of the two groups (so long as there are no errors in our social-status scale systematically associated with differential preferences for certain occupations between the groups);

(2) will provide a more precise comparison, so long as the "slope" used is somewhere near that slope which would be most effective (there is no necessity to get the slope "exactly right", though better slopes will be more effective);

(3) will not compensate for the curvilinear part of the relationship between the measure of social status used and average income.

In view of (1) and (2) we have made considerable progress by using adjusted income. In view of (3) we may well wish to tabulate mean adjusted incomes by social class, thus still further freeing our comparisons from the effects of social status.

This example has on the one hand been simple, and on the other nonexplicit. It is hoped that its simplicity will outweigh its nonexplicitness and that it will throw useful light on the possibilities of "adjustment." These possibilities are many, varied, and rewarding. Skill in their use comes from practice, from thinking about just what is being done (rather than staring at formulas), and, above all, from common sense.

(The one example of "adjustment" frequent in the behavioral sciences is the adjustment of death rates for the age distribution, which is far more a matter of "standardization" than of "adjustment." It is a useful procedure, but cuts nowhere nearly as deeply as most adjustment procedures.)

F4. THE USE OF RESIDUALS

A key part of almost any technique of several-stage analysis is the use of residuals. This differs in purpose from the use of adjusted values, though not necessarily in computation. It is merely a matter of definition whether residuals are the result of adjusting the observed responses for the effects of all variables which are naturally, or reasonably, considered, or are the deviations of observed responses from fitted responses, i.e., from those values predicted on the basis of all variables which are naturally, or reasonably, considered. In either event, residuals represent the deviation of what is observed from what has been systematically described. Whereas adjusted values are intended for use in planned further comparisons, residuals are calculated either (i) as a basis for estimating the size of fluctuation (as a basis for an "error term") or (ii) as a step in discovering *unsuspected* phenomena. The former purpose is more classical, but the latter is probably more important.

The use of residuals is an art where some physical scientists long maintained a significant lead on most, if not all, statisticians. What are considered to be the "raw" results from many a physical science experiment are not the responses themselves, but the residuals, the differences between what was observed and what currently-used theory "predicted". This fact has been one of the important keys to physical science progress.

Many of the more powerful forms of statistical analysis developed since 1920 can be formulated in terms of residuals. Almost all analyses of variance can be regarded as making use of residuals. In a two-way classification, for example, the formal numerical "interactions", the sum of whose squares is the "interaction sum of squares", may be regarded as the residuals after fitting both row and column main effects. For a long time the analysis of variance was used in such a way as to miss real opportunities of discovery. Residuals were used only as an error term. Indeed, individual residuals were almost never calculated, only the sum of their squares (which certain algebraic identities usually make relatively easily available) being obtained. In the last few years, however, because of the recognition that much could sometimes be learned form the values of individual residuals, and because of the increasing cheapness of routine calculations, the calculation of individual residuals has become much more frequent. This trend will continue.

F5. ADJUSTMENT AND RESIDUALS FOR COUNTED DATA

In those situations where the raw datums of behavioral science are of an irretrievably "yes-no" nature, it is clear that we cannot very well adjust individual observations, or find useful residuals corresponding to them. But this fact is not the insurmountable bar to the use of such modes of analysis that it might seem. For in such circumstances the relevant unit of data is not likely to be the individual observation.

In these situations, the relevant unit of data is much more likely to be a small (or larger) group of individuals, and its numerical content is likely to be expressed by the counted fraction of individuals who are "yes" or "no". Whatever mode of expression for this counted fraction may serve us best, percentage, anglit, normit, logit, or "what-have-we", will be such that meaningful residuals can be calculated, that meaningful adjustments can be made. (Indeed, if we are wise, we shall frequently choose the mode of expression to maximize the usefulness of these residuals or adjusted values.)

(It is possible to carry on a somewhat related procedure with individual "yes-no" observations. Somewhat more esoteric techniques, including the fitting of logit planes and the use of Fisher's scores, then enter naturally. Since many applications to behavioral science of simpler techniques are applications at present relatively untouched, we need not try to discuss these more sophisticated techniques here.)

F6. ANALYSIS OF DATA IN GENERAL

The moral of our whole discussion of the analysis of data is that, while it is not simple, it can be very rewarding. The extent of the nonsimplicity can usually be adjusted to the skill of the analyst, and to the pressure of time. Such adjustments are necessary, but a little more effort often provides substantial returns.

The greatest mistake in approaching data is often the idea that analysis of data is like taking the hook out of the fish's mouth, to be done once and for all, and as expeditiously as possible. There are many ways in which it pays to analyze data in stages, and many more will be discovered.

Heavy emphasis needs to be given to the advantages of pre-analysis of a small sample of the data. A few years ago, friends and colleagues of mine collected, with considerable effort, 1000 long questionnaires; they punched up cards, hired a sorter, and started cross-tabulation. By the time they had reached their limits on time and

funds, they had come to see fairly well what was really going on, to recognize which of the tabulations and analyses that they had *not* made would really illuminate the social situation. The clock having struck, these tabulations could not be made, and useful information and insight had to be left buried in the data.

How often does this happen? (And are our practices with the timing of data-collection and thesis-submission such as to educate each new Ph.D. to regard this as the usual thing?)

Now in this particular instance, as I came to see far too late, simple actions could have avoided a large part of this difficulty. The field work took a long time, and was mainly in the hands of interviewers other than the chief investigators. A random subsample of 10 questionnaires could have been available a month or more before the end of the field work. The investigators could have sat around a table for several days, chewing over all the details of these 10 questionnaires, all the patterns they suggested, what analyses would reveal these patterns. (And by this process, not only the information in these 10 questionnaires, but also many of the insights in the investigators' minds, would have been mobilized and made available to guide analysis, instead of being dredged out at the last minute to explain why Table so-and-so might reasonably behave in the otherwise peculiar way that it does.)

Then they could have expanded this subsample to 50 or 100 cases and *written* the parts of the data which seemed most likely to be relevant on ordinary cards (edge-marked if desired). There would then not have been too many cards for any kind of hand sorting and hand tabulation. The suggestions of the 10-questionnaire subsample, and any other available suggestions could have been tried out. And from more discussion there could well have come more insight. At this point, for the first time, a rational plan for the analysis of the whole sample could have been developed, wherein it would be possible to specify not only which tabulations seemed necessary, but also which were needed *first*. (A tabulation may be needed early (i) to provide an opportunity for the facts to force reconsideration of current suppositions, or (ii) to help in choosing alternative continuations of the analysis.) By this time the 1000 questionnaires would have been almost available.

Had this sort of a step-by-step approach been adopted, and the same amount of time and money given over to the analysis (including pre-analysis), I am convinced that much more would have been learned. How often is this the case? It seems that it must happen "all the time."

The advantages of pre-analysis have been noticed by behavioral scientists, at least in footnotes. Thus Hyman (1955, p. 332-333 footnote) says: "a valuable procedure .... is that of trial tabulations, in which ideally a random sample ..... is processed first. This ..... yields quick and generally reliable estimates of the larger findings."

It would be a mistake to believe that such step-by-step analysis is natural only for the exploratory inquiries associated with (i) a long questionnaire, (ii) depth interviewing, or (iii) certain aspects of social anthropology. A certain psychological experiment, in which the behavior of subjects under quantitatively-described conditions was quantitatively measured, appears in the literature as a methodological example (Johnson and Tsao 1945, Johnson 1949). Its purpose was to illustrate the use of complex analyses of variance in such connections. It makes use of randomization, balancing, replication, factorial design; of most of what have been called (Student 1938, p. 365) "all the principles of allowed witchcraft." And a "corresponding" analysis of variance was published, which we will label "Analysis 0." Unfortunately, this analysis was clearly inappropriate. So a colleague and I (Green and Tukey 1960) made a new analysis, "Analysis 1", which avoided certain inadequacies and could be extended to a more complete and appropriate analysis, "Analysis 2". So long as the actual numbers were not examined carefully (i.e., with the aid of an adequate analysis) there could be no objection to Analysis 2. In principle, indeed, it *could* have been appropriate.

But when the numbers were looked at, certain unsuspected relations among them clearly indicated a different approach. So we went back and made "Analysis 3", which was sensitive enough to uncover new regularities and led to "Analysis 4" (not yet published). This last analysis was quite self-compatible (the only indications being that if the actual randomized orders in which the subjects underwent various situations were known, some further gain might be made by adjustment for learning within a session). The answers, which were now rather sharply defined, were not very close to those of Analysis 0, or even to those of Analysis 2. The truth was *in* the original 448 observed numbers, but it took several stages of analysis to bring it *out*.

The moral is clear. Analysis by stages may be necessary with any data, even when gathered in the pattern of a formal, randomized, "experimental design".

G. SIGNIFICANCE AND CONFIDENCE

Many questions involving significance testing produce active debate. What purposes can and should significance procedures serve? Should routine significance testing be used in exploratory sociological research? What needs to be done to avoid the common fallacies of significance

testing? How can we modify significance tests to make them more relevant to our problems? To what extent should significance procedures be replaced by confidence procedures? How should we apply significance or confidence procedures to (a) multiple comparisons, (b) the results of complex calculations?

Full answers to all these questions would require a few books. Only indications can or should be given here. But even simple indications, if taken up actively, can lead to substantial improvement in the day-to-day handling of data.

We have already discussed why significance tests cannot settle causality (in A) and how the choice of significance procedures should be related to the intended length of inference and the choice of hypothetical population (in B). It is now time to go on to some of the other questions.

G1. WHEN SHOULD SIGNIFICANCE TESTS BE USED?

There is much discussion of when significance tests should be used in sociological inquiries. In most such discussions, "significance tests" are equated to "classical tests for the significance of the difference of fractions based upon assumed independence of sampling of individuals". Perhaps the most important issues are reflected in such brash statements as:

(1) Significance tests don't establish causality, and we aren't interested in anything else. (See A above.)

(2) The classical significance tests are inappropriate because their presuppositions do not hold. (See B above for some general considerations, Kish 1957 for more specific difficulties which occur even in efficient probability samples, and H4 to H6 below for a practical way out of many difficulties.)

(3) It isn't SCIENCE unless you prove it; beside mathematical proof, and proof by experimental manipulation, only sanctification by significance tests constitutes proof; sanctify or die!

(4) If I put in some significance tests to sanctify my results, no one can complain about anything, not even about those weak techniques.

(5) Much valuable work in any field is exploratory; exploratory work must seek out even feeble indications; exploratory work dare not throw away indications just because they are not significant at some conventional level.

(6) Admittedly there must be exploratory work, but is it sensible to write books and books about its results?

(7) Even in physics and chemistry, the ultimate standard is repeatability by different workers at different times and places and each worker publishes his own work. Why should there not be a whole book about each exploration? After all, agreement of enough explorations will produce very high quality significance!

The case for exploration untrammeled by statistical significance is put forcefully by Zeisel, who says: (Zeisel 1955)

> "There is, now, in the social sciences no greater need than the development of theoretical insights guided by empirical data — to provide this guidance and serve as stimulant — [we rely on] the significance of statistically insignificant data. Even if the probability is great that an inference will have to be rejected later, the practical risk of airing it is small. Subsequent and more elaborate studies may disprove some of these inferences; but for those that survive social science will be the richer."
>
> "To be sure, a physicist would rightly frown on such recommendation. But his is a world with generalizations on a high level. By comparison, the social sciences are at a stage where for decades to come the formation of even tentative theoretical structures will be at a premium."

(See also Merton et al. 1957, pp. 302-303.)

It is not for a statistician to lay down a schedule of dates for the change in emphasis from exploration to confirmation in each of the many subsubfields of behavioral science. It is right, however, to lay a heavy charge upon the consciences of all individual behavioral scientists, and upon the collective scientific consciences of each of the varied fields and subfields of behavioral science, that

(1) there are appropriate places for both exploratory and confirmatory work;

(2) repeating unconfirmed work is useful, perhaps necessary, not wasteful;

(3) significance based upon repetition under suitably diverse circumstances is essential to confirmation;

(4) it is very hard to justify holding back exploration with severe requirements of statistical significance.

This is another place where simple answers should not be forthcoming, a place where each must help to bear the "quantitative man's burden".

G2. *THE SEARCH FOR CERTAINTY*

Many behavioral scientists who use statistical techniques are novices as statisticians. As in so many other instances, "ontogeny repeats phylogeny" in the nature of the help they think it appropriate to seek from statistics. Just as Reichenbach (1951, e.g., p. 117) portrayed the development of philosophy as a learning that more and more questions should not be asked, so the development of statistics can be portrayed as learning of more and more things about which certainty should not be sought. A brief, oversimplified outline of the development of statistics through the last half dozen decades may help to illustrate the point.

The first real step toward modern statistics was the work of Karl Pearson. Much of his impact can be interpreted as a shift from an implicit certainty that samples matched the populations from which they came to a certainty that random samples did this closely enough if they were large enough.

The impact of Student (William Sealy Gossett) was in large part the recognition that, starting with a small random sample, even if you cannot be certain about the population, you can be certain about the degree of uncertainty of your inference. (In particular, you can be sure about your level of significance.)

R. A. Fisher (now Sir Ronald) extended the implementation of exact tests of significance in many directions, and introduced very important methods of dissection of data (such as the analysis of variance), but from the aspect we are now considering, his greatest impact was through an attempt to restore more certainty to inference. Admittedly the result of your inference from a small sample could not be free of allowance for sampling fluctuation, but you could try to be certain that the inference procedure you used was the best possible, often because it "used all the information." (On Student's urging (cp. Pearson 1939, pp. 242-243), a modified approach to this kind of certainty was pioneered by J. Neyman and E. S. Pearson.) The more classical kind of statistical certainty was extended by the growth of a wide variety of interval estimates from the twin skeins of the first fiducial (Fisher) and confidence (Neyman) statements.

Then the work of Abraham Wald rolled back certainty still further, when he showed that, insofar as procedures leading to definite actions were concerned, there could, in general, be no single optimal statistical procedure, but only a "complete class" of procedures, among which selection must be guided by judgment or outside information.

All of this development made heavy use of closely specified population models in which only a few parameters were left free to match the model to the real world. During the next decade or decades, the growing recognition of the unrealism of such presuppositions will inevitably drive statisticians to less certainty about the optimality of procedures, to greater reliance on experience in and near particular fields of application.

This summary is undoubtedly somewhat unfair to individual fields of statistics and to individual statisticians. But insofar as it presents the growth and burgeoning of statistics as involving a relentless continued pressure for certainty, and a forced abandonment of one certainty after another (til the writer tries only to be certain that one cannot be certain), it is ultimately truthful and deeply revealing.

There should be no surprise that the statistical novice seeks for certainty in statistics. There is no ground for anything but gentleness in readjusting his goals. ("He needs help.") But there is equally no warrant for leaving them unadjusted.

G3. SOURCES OF UNCERTAINTY

What are some of the reasons why certainty cannot be obtained by statistics? Not because of small samples, because Student and Fisher have shown us how to be certain about uncertainty. In part because our models will always be approximate reflections of the real world. But especially because not all sources of variability have had a chance to show their effects by entering differently into two or more parts of the data before us.

Some sources of variability have not been revealed because the data had no chance to come from the whole range of possibilities with which we are concerned. There are many reasons why the sampled population never coincides with the target population. (The difference in epoch between observation and interest, which applies to all except the purely descriptive historian, is but one reason for non-coincidence.)

The nature of measurement is also concerned. Not just the fact that measurement is fallible, subject to fluctuations. For fluctuations which are seemingly random, or which are associated with small changes in time or space, will be represented by differences between parts of the body of data before us. Physical scientists have been keenly aware for a very long time that there are systematic errors in *all* their measurements. These vary from errors intrinsic in the definition of the measuring instrument (which cause that which is measured to differ from that which is said to be measured), through errors intrinsic in a

specific realization of a measuring instrument (which make the average readings of one instrument disagree with those of another), to fluctuations which are associated with large changes of time or space, and which were consequently not explored, by either the actual data or by the potential data, as identified in the sampled population.

The physical sciences live with these difficulties and uncertainties every day, are continually conscious of their existence, are aware that there are troubles with which statistics cannot appreciably help them to cope. Physical scientists neither deny the existence of these difficulties, nor give up because they exist. Physical scientists try to use statistics, be it formal or informal, to deal with *those aspects* of their problems to which statistics is suited. (The fact that some branches have not yet begun to use certain of the newer techniques is an evidence of ignorance of fact, not a misdirection of philosophy.) Can behavioral sciences and behavioral scientists do less?

It is in the face of such inescapable uncertainties that we must use statistics. How then can we allow the mildly uncertain character of a hypothetical population (see B3) to worry us seriously? How can we flee to statistics as a source of certainty, as a way to avoid all our difficulties and troubles?

To some, of course, statistics is a refuge, not from the doubts and fears of the individual investigator himself, but rather a refuge from the criticism of colleagues, readers, and listeners. "If only my results could be sanctified by statistics, my techniques and selection of material made immune to criticism, how wonderful it would be!" Many users of statistics seem to feel this way. It is not surprising that they do. But such an attitude can only retard the progress of science; if sufficiently widespread it could stunt the growth of science or even deform it like a Japanese miniature tree. We dare not let statistics be a *general sanctification.* It can rightly offer evidence as to the uncertainties due to finiteness of data, and offer, in *experimental* situations, evidence about many aspects of the adequacy of controls and comparison (Fisher 1935ff, p. 2). It is badly needed in these limited roles, but it must be kept in its place.

G4. FALLACIES OF SIGNIFICANCE TESTING

Two contradictory sorts of false optimism tempt us whenever we assess the usefulness of a tool, even of a statistical tool, on the basis of how it is used. We may be falsely optimistic about tools and tool users and, believing that it is possible to make tools which will always, or even nearly always, be used correctly: we may decry all tools in

proportion to the number of times each is misused. Good tools are then likely to be decried much more vigorously than bad ones. For they will be used so much more frequently that even small percentages of misuse will lead to large absolute numbers.

On the other hand, one could be falsely optimistic about the responsibilities of our roles as tool purveyors (whether as tool designers, tool salesmen, or tool advisors). For one might (as too many do) feel that a tool is well enough developed if it will be properly used *when* used according to all the rules on the package, including the fine print. Such views disregard the empirical observation that no tool is so used.

The ridge between these valleys of misplaced optimism is narrow, yet we all should traverse certain parts of it. In describing some of the commoner fallacies of significance testing, it is not our purpose (as it seems to have been in Selvin 1957) to assert that significance testing is intrinsically evil, for that is not at all our view. Rather we are trying to point out some of the places where improvements in the use of significance testing are both possible and desirable, in the interests of making the tool more valuable.

It is natural, when picking up a new tool, to hope, and perhaps even to think, that it will do for you just what you wish it to do. The classical fallacy of significance testing, is badmandment #100, namely:

The significance level tells you the
probability that your result is WRONG.

Every statistician spots some form of this badmandment quite frequently, and marks up its appearance as an error. The statistician is logically (and interpersonally) right in this ascription, for the significance level does not provide a probability of the result being wrong, as simple examples show. Instead it provides something notably different (which is described clearly in many textbooks). Reproducing such statements is interpersonally wrong, because some readers may know just enough to mislead themselves. Intra-personally, however, where the investigator as data analyst is trying to communicate with himself in his other role as interpreter of the results of data analysis, such phrases may not be too dangerous. Investigators who use them may not be clear enough during *intra*-personal communication, and may not need to be clear enough, about "probability" or "the probability of being wrong", to mislead themselves seriously. Investigators who do use such phrases do seem to interpret data in about the same way as those who do not. Intra-personally, then, this fallacy is not too serious, although, especially for the benefit of younger readers, it ought clearly to be kept out of the printed literature.

Significance testing is the subject of many serious fallacies; an exhaustive list would be out of place. Examples of one family are discussed below. Certain others are implicitly covered by the discussion of the next two sections, while still others can be found in textbooks and in the parts of Selvin's article (Selvin 1957) which are not concerned with the failure of significance tests to establish causation.

Leading our short list is perhaps the simplest fallacy of them all, an inevitable consequence of trying to render a portrait with a single round dot, either black or white, (of trying to send a one-bit message). Suppose that Jones and Smith, separately, do the same experiment, measure the same difference, and make tests of significance. One verdict may be "significant" the other "not significant". Then Robinson writes a review paper stating that Smith and Jones found contradictory results. But a confidence analysis may show, and usually does show, that *both* experiments are consistent with any one of quite a wide range of values for the differences concerned. Clearly more definiteness is being read into the statements of bare significance procedures than belongs there.

In a parallel fallacy, a single experimenter or observer examines the effects of two variables, finding one "significant" and the other "nonsignificant", concludes that the first is more important than the second, although there may be far from enough evidence to show this.

In a third example, an experimenter may measure some of the more interesting of the 561 correlation coefficients among some 34 variables. Perhaps 25 out of the 30 which he regards as interesting turn out to be "significant" (i.e., to have been successfully discriminated from zero). He is dangerously likely to use this judgment of significance not merely to convince himself that the relationship indicated by a correlation coefficient which just barely reaches his chosen level of significance is stronger than that indicated by a coefficient which just fails to attain this level, but even to support a belief that the order of relative size among "significant" coefficients is exactly (or, at least, *almost* exactly) the same as the order of size of the population coefficients he is estimating. Such data may not contain any appreciable evidence in support of any of these views.

These particular fallacies of significance testing come from failure to recognize that a classification into only two classes is not necessarily a classification into clearly defined classes, into classes that are "broad" in the sense that classification is easy and reproducible. The classes "statistically significant" and "not statistically significant" are *narrow* in the sense that independent reclassification, namely repetition of the whole experiment or observation on independent chosen individuals or under independently chosen circumstances would differ from the

original classification in a non-negligible fraction of all instances. (Changing to a still broader classification is no help at all in this instance.)

A large measure of the situation is summed up in Yates's 1951 sentences: "Usually quantitative estimates and fiducial limits are required. Tests of significance are preliminary or ancillary." Many users of statistics have learned to replace significance techniques by confidence techniques; more should and will.

H. TECHNIQUES OF SIGNIFICANCE AND CONFIDENCE

Principles of significance are important, but they gain their value by being combined with techniques. This account has no place for the many significance (and confidence) techniques which are well described in available books, but it can and should summarize the most useful ones which are not easily available.

Basic to such a discussion is a clearer understanding of the notions of significance level, and of its generalizations and revisions. The basic notion is of an accepted chance of being wrong. There is need to describe and specify:

(1) What it means to be wrong;

(2) What chance is accepted;

(3) Under what circumstances this chance must be faced.

Even in the simplest case, say a two-sided t-test at the 5% level against equality of means, there are two interpretations. One is:

(1) to be wrong = to assert statistical significance when the population means are in fact equal;

(2) the accepted chance is 1 in 20 such comparisons; and

(3) the chance is only faced when the two population means happen to be equal.

This first three-part interpretation applies when "statistical significance" is taken to mean "are not equal." Another interpretation is:

(1) to be wrong = to assert statistical significance in one direction when the population means do not differ in that direction;

(2) the accepted chance is 1 in 20 such comparisons; and

(3) the whole chance is accepted when the two population means are equal; a lesser chance is accepted when the population means are close, but not equal; this lesser chance falls to zero as the separation between population means becomes indefinitely large.

This latter three-part interpretation applies when "statistical significance" is taken to mean "one population mean differs from the other in the same direction that the one sample mean differs from the other".

It is convenient to call the *largest* chance of error that has been accepted the *error rate*. This term emphasizes that such a quantity is a fraction whose numerator is a number of errors, and whose denominator is a number of chances to make an error. Under different circumstances it will prove wise to use different definitions of what is "one" error and what is "one" opportunity to make an error.

An investigator who works to a nonextreme significance level, such as 5%, and who only compares means which are easy to distinguish, perhaps using large samples from populations with widely different means, will rarely make an error in asserting statistical significance. All of his differences *are* really not zero, and he has made it easy for the data to *show* that they are not zero. It is very hard for him ever to be wrong. He has many formal opportunities to make errors, but the actual chance that any individual opportunity will produce an error is small. He has budgeted a 5% error rate, but he is not really spending it.

Errors are "bad things," so that it is quite natural to compliment him on not spending his error rate. Closer analysis shows, however, that he deserves no compliments. Suppose his long lost identical twin is making exactly the same studies, but was taught to use a 1% error rate. The latter sib will reach almost exactly the same conclusions, but he will reach them at 1% rather than 5%, attaining greater security from the same data. There is a loss in not using the error rate that has been budgeted, as can also be seen by considering many sorts of situations which differ markedly from the one just described.

It is a truism easier forgotten than remembered that, just as any sample can come from any normal distribution (though much more probably from some than from others), so any body of experimental or observational data can come from almost any underlying situation (though much more probably from some than from others). Almost all empirical knowledge is purchased at the price of error. Error rate is one of the coins that is paid for knowledge.

Once we recognize error rate as a medium of exchange to be budgeted, we must be prepared to divide its application to various ends, like anything else to be budgeted, in the way that yields us the highest return. The next two sections will provide an example of how this may be done.

Once we recognize that not spending the error rate that has been budgeted may be wasteful, we seek to find out how to spend it more completely. One way is to change from a significance procedure to a confidence procedure: to assert, for example, when appropriate, with 95% confidence that the difference between the mean for *A*'s and *B*'s lies between +2.1 and +7.3, instead of asserting either that the difference is positive at the 5% level, or that the difference is significantly not zero at the 5% level.

In particular, this change has the great advantage that it makes it possible to distinguish "negative" results of very different strengths. To have 95% confidence that a difference is between −4.3 and +11.2 is quite a different thing from having the same confidence that the same difference lies between −0.0032 and +0.0017. Yet either is properly reported in significance terms as "the observed difference was not significant at the 5% level".

Most significance procedures have directly corresponding confidence procedures. It is reasonable to argue that the presuppositions of a corresponding confidence procedure are somewhat less likely to hold than those for the significance procedure. (Thus, for example, assumption of similar shapes and variances of two distributions to be compared is more reasonable when testing whether the means of the two distributions may be the same, than is the same assumption when testing whether the difference between the two means may be −14,329.) But wide experience suggests that the more efficient use of the budgeted error rate far outweighs such considerations. Confidence procedures for simple situations are widely spread through the literature. One aspect of confidence technique deserves especial attention here, however. Multiple comparisons, in which at least all pairs of 3 or more means (or slopes, or what have you) are to be compared, arise frequently. Since there is little in the literature on this topic, Section H3 offers a brief discussion, in which the nature of an error rate as a fraction plays a central role.

Finally, there is the problem of the results of complex calculations. Even if the relative numbers of investigators and statisticians were reversed, the few investigators could produce new ways to combine data faster than the many statisticians could develop and package appropriate specialized significance or confidence procedures for these new combinations. There is a great need for a nearly universal technique,

which must be easy enough to use, but need not be perfect in any respect. Like a Boy Scout jackknife, such a technique should be usable for anything, although, again like a jackknife, each of its jobs could be better done by the corresponding specialized tool, if that tool were only at hand. The last three sections (H4 to H6) are an introduction to such a tool.

H1. SPLITTING AND ALLOCATION OF ERROR RATES

The idea of the single overall test of significance as something natural, universal, and perhaps even as a cure-all, might almost be considered a statistical disease. As such it is "panstatistic" rather than "epistatistic," being present almost anywhere and over a long time.

In the statistics of measurement it most naturally and frequently appears in the use of a single overall F-test as the comparison among all means. The development of the analysis of variance has provided, as one of its implicit, unrecognized functions, a way to escape from such an overall test. (These tests are sometimes called "portmanteau tests" because they try to carry everything at once. Perhaps "carpetbag test" would be more appropriately degrading.)

As a matter of fact, however, the classical approaches and practices of the analysis of variance involved concealed inequalities in the way in which error rates were granted to different blocks of intercomparisons. The development of multiple comparison procedures during the present decade (see H3 below) has thrown light on these inequalities, and developed a certain conflict of opinion, and considerable intensity of view, as to whether they should be removed by weakening the more stringent standards or tightening the less stringent ones. (The writer has upheld the latter view.)

Insofar as the analysis of counted data is concerned, the F-test and the analysis of variance have been historically of limited significance (though this situation is likely to change over the years). Thus any example we give here should relate to some other portmanteau test and its modification. The grand portmanteau for counted data is of course the chi-square test, and it is here that we shall deal with an example. But before coming to the example itself we need to clarify certain general aspects.

First, as we noted above, there is not yet unanimity as to what should be the customary assignment of error rates to dissected F-tests (or their replacements). And it is almost certain that whatever view may prevail for F-tests, an analogous view will in due course prevail for chi-square. The writer is a protagonist of one side in this discussion; will

those who follow his views thereby put themselves in jeopardy? I feel that we can confidently say "no". For the procedures to be discussed are the most conservative of all the multiple comparisons procedures so far proposed, in the sense that every indication they call "definite" will be called definite at the same probability level (or perhaps at a more extreme level) by any other multiple comparisons procedure. Yet the procedure to be discussed will be seen in practice to be *more* sensitive and searching than the classical overall procedures. However the controversy comes out, we can make a definite gain in sensitivity and incisiveness by taking this step now.

Second, empirical knowledge has to be bought by the simultaneous expenditure of several currencies. One of these is a willingness to be wrong a certain fraction of the time. We spend this currency, along with others, whenever we do an experiment or make an inquiry. We can spend varying amounts of "error rate" at our choice, and it is best to spend the most where the probable return is most valuable. Quantity and price are far from linearly related, however, and we may expect to be wise by spending many small sums upon individually unpromising possibilities, together with fewer larger sums upon individually promising possibilities. Thus we should allocate, or budget, error rate with the same care with which we allocate samples over strata, or numbers of levels over factors, and, in this allocation, we should be guided (in part) by some of the same principles which guide the other allocations.

H2. A SPECIFIC EXAMPLE

The specific example to be dealt with comes originally from a biochemical study of people (Williams, et al. 1950), in which some 62 significance tests were performed. These tests seemed on the whole to show association, though no one, two or three gave conclusive evidence. The initial treatment was objected to (Popham 1953), so that a further, more formal analysis was undertaken (Tukey 1954). (For further discussion of the original example, see Chung and Fraser 1958 and Dempster 1960.) As part of an early stage of this further analysis, it was desired to treat these 62 significance tests as if they were independent. (Strictly speaking, of course, they could not be independent since they involved the same persons.) In the original analysis, two-decimal two-tailed P values had been obtained for all 62 significance tests. The problem before us is thus: Is it reasonable that these 62 values are a sample from a uniform distribution? To answer such a question we naturally reach for a chi-square test. (Before going on to the details, we

should remark that it is possible to make suitable allowance for the actual lack of independence of the 62 P values, and that this was also done in Tukey 1954.)

The conventional chi-square analysis is set forth in Table 18, where an unusual column is to be found at the right. The conventional result, then, is a chi-square of 14.47 on 9 degrees of freedom, corresponding to a significance level of about 11%. Those who use this analysis have behaved as though it were equally valuable to detect any and all deviations from uniformity of distribution. This is of course far from the case. Deviations of rather systematic natures, corresponding to the piling up of P-values at one or both ends, or in the middle, are *both* more likely to occur *and* more valuable to detect. We should in some way focus more of our attention (i.e., spend more of our precious error rate) on deviations of these sorts, leaving less for more irregular deviations from uniformity.

Having our chi-square written as a sum of squares, we may apply conventional techniques of breaking up sums of squares and examine the results appropriately. When the elements contributing to a sum of squares are naturally arranged at equal spacing in a one-way table, as is the case here, it is natural to use orthogonal polynomial coefficients as a means of dissection. The gory details, which need not really concern us, are to be found in Table 19. The results, with which we should be concerned, are given in Table 20. The right-most column labeled "Allotted 'bogey'" is again unfamiliar. Its contents, 2%, 2% and 1%, represent a splitting up of the 5% error rate we would otherwise have been willing to allot to the whole 9-degree-of-freedom chi-square, and an allotment of these parts to the three portions into which this chi-square has been split. The choice of the sizes of these three parts of 5% has obviously to be a matter of judgment, but the realities of the situation go far to prescribe what should be done.

In the actual example, the trend constituent, with its significance level of 0.2%, is far more extreme than the allotted bogey of 2%. Consequently we conclude that the P-values are piled up at one end or the other. We may if we choose, and it would be very generally wise to so choose, allot this 2% half to the trends toward high P-values and half to trends toward low P-values. The observed one-sided significance level of 0.1% exceeds the new one-sided bogey of 1%, and examination of Table 21 shows that we may conclude that there was an excess of *large* P-values.

Table 18

Application of classical chi-square to the comparison of 62 P-values with a uniform distribution from which they might have been a sample. In each cell, O = number observed and E = number expected. Value marked * is χ^2_9 as classically calculated.

Cell boundaries	O	E	$\frac{(O-E)^2}{E}$	$\frac{O-E}{\sqrt{E}} = x_i$
0.00-0.09	3	6.2	1.65	−1.29
0.10-0.19	2	6.2	2.84	−1.69
0.20-0.29	5	6.2	.23	−0.48
0.30-0.39	5	6.2	.23	−0.48
0.40-0.49	8	6.2	.52	−0.72
0.50-0.59	7	6.2	.10	−0.32
0.60-0.69	3	6.2	1.65	−1.29
0.70-0.79	8	6.2	.52	−0.72
0.80-0.89	9	6.2	1.26	−1.12
0.90-0.99	12	6.2	5.44	−2.33
Total	62	62.0	14.44*	−0.02

χ^2_9 also = $(-1.29)^2 + (1.69)^2 + \cdots + (-2.33)^2 = 14.47$ and very nearly = $(-1.29)^2 + (1.69)^2 \cdots - (2.33)^2 - (0.02)^2 = 14.47$

* Differs from 14.47 due to accumulated roundings.

As was the case in this example, such dissected tests, which spend no whit more of error rate, often both:

(1) detect deviations which would be otherwise unnoticeable, and

(2) report their positive findings in much more specific and useful forms.

Such procedures of splitting and alloting error rates deserve consideration, and use, in a wide variety of situations, including complex analyses of variance. Detailed discussion here, however, would lead us too far afield.

Table 19

Breakup of the 9-degree of freedom sum of squares which approximates classical chi-square into three parts.
$x_i = (O-E)/\sqrt{E}$ for i^{th} cell. SSq = sum of squares for column.

Cell	x_1	Orthogonal coefficients		Products = (x_i) × (coefficients)	
0	−1.29	−9	6	11.61	−7.74
1	−1.69	−7	2	11.83	−3.38
2	−0.48	−5	−1	2.40	0.48
3	−0.48	−3	−3	1.44	1.44
4	−0.72	−1	−4	−0.72	−2.88
5	−0.32	1	−4	0.32	−1.28
6	−1.29	3	−3	−3.87	3.87
7	−0.72	5	−1	3.60	−0.72
8	−1.12	7	2	7.84	2.24
9	−2.33	9	6	20.97	13.98
S (= sum)	−0.02	0	0	−55.42	−6.01
SSq	14.47	330	132		

$$\chi^2_{linear} = +(55.42)^2/(330) = 9.31 = (+3.06)^2$$
$$\chi^2_{quadratic} = +(6.01)^2/(132) = 0.27 = (-0.52)^2$$
$$\chi^2_{residual} = +(14.47) - (9.31) - (0.27) = 4.89$$

Table 20

The dissected chi-squares and their relation to "bogey".

Nature of chi-squares	Degrees of freedom	Chi-Square value	Significance level	Allotted "bogey"	
Indicative of piling up at one end or the other	1	9.31	0.2%	<<	2%
Indicative of ends higher or lower than the middle	1	0.27	60%	>>	2%
Residual	7	4.89	55%	>>	1%
		(Undissected for comparison)			
Pooled	9	14.47%	11%	>	5%

Table 21

Further dissection of first single degree of freedom in Table 19

	Deviate	Deviate value	One-sided significance level*		Allotted "bogey"
Indicative of piling up toward cell 0	$-\chi_{Linear}$	−3.06	99.9%	>>>	1%
Indicative of piling up toward cell 9	$+\chi_{Linear}$	+3.06	0.1%	<<	1%

H3. MULTIPLE COMPARISON PROCEDURES

As noted in passing above, the present decade has seen the development of multiple comparison procedures, ways of intercomparing a number of means or other estimates in all possible ways. Specific techniques have been proposed by a number of authors. (For some references see Kurtz et al. 1965a.) The most conservative of the serious proposals is that of the writer, which can be summarized as in Table 22. The necessary factors are provided in Table 23.

Through the courtesy of Frank Beach, we can present an example which is surely behavioral, since it involves both sex and rats. It is adapted from an unpublished study by Beach and Whalen on the effect of enforced rest, following one ejaculation, on copulatory behavior in rats. The variable treated here is the average intercopulatory interval (ICI) after the enforced separation of the rats. The data and computations are presented in Table 24. Using this technique, it is, in particular, demonstrable with 95% confidence that the ICI's for 15 and 60 minutes are *each* less than those for *either* 5 or 180 minutes.

H4. THE BASIC SOURCE OF CONFIDENCE

It is time to say again, specifically, firmly, and clearly, what we have said before: The only basis for statistical confidence, including confidence in the statistical significance of some difference, is the presupposed independence of the fluctuations contributed by different portions of the body of data considered. It may well be that this restriction is not necessary, it may be that other presuppositions might come to serve as alternative bases. But they have as yet not done so.

Table 22

Digest of a multiple comparisons procedure.

If $x_1, x_2, \ldots, x_m$ and s^2 and f are such that, *had* any x_i and x_j been the *only two* x's, it would have been legitimate to refer

$$t = \frac{x_i - x_j}{s\sqrt{2}}$$

to Student's t distribution on f degrees of freedom so that, *under that assumption,* we could (a) test the significance of the difference by comparing this value of t with the value $t_{5\%}$ found in the standard tables for 5% and f degrees of freedom; (b) assert with 95% confidence that the true difference differed from that observed by no more than $\pm t_{5\%}(s\sqrt{2})$; then, *under the actual conditions,* we need only increase this comparative allowance by multiplication by a factor which may be taken from Table 23, forming

$$\text{comparative allowance} = \pm (t_{5\%})(s\sqrt{2})\,(\text{factor}).$$

We may then assert any and all observed differences $x_i - x_j$ as differing from the corresponding population (or "true") difference by no more than this comparative allowance, and that *all* of our assertions will be correct in 95% of all instances. (One instance — application to the intercomparison of *all* pairs is one family, $x_1, x_2, \ldots, x_m$, of determinations.)

(There is no mathematical reason why a presupposition that certain fluctuations are not independent, but instead have all simple correlation coefficients equal to 0.57, could not serve as such a basis. But the reasons discussed in A2, which are mainly psychological in nature, do a lot to cause such alternative presuppositions to be regarded as arbitrary and to prevent their acceptance in practice.)

Thus in simple random sampling, for instance, we presuppose that each item is selected at random from the population, a presupposition which ensures correct results "on the average" but provides no basis for significance or confidence, *and* that these selections are independent, which leads to significance tests and confidence procedures which are valid whenever the appropriate, detailed, additional presuppositions hold. Selection at random of clusters of items, rather than single items, has no effect upon the correctness "on the average" of the sample result, but greatly changes the appropriate technique for judging confidence or significance. Indeed, if the clusters cannot be even partially identified in the given data, there will be no way to attain statistical confidence and significance.

Table 23

Factors for calculating comparative allowances.

m	factor	m	factor	m	factor	m	factor
		11	$1.65+1.7/f$	21	$1.82+2.5/f$	35	$1.95+3.4/f$
2	$1.00+0.0/f$	12	$1.67+1.8/f$	22	$1.83+2.6/f$	40	$1.99+3.7/f$
3	$1.20+0.4/f$	13	$1.70+1.9/f$	23	$1.85+2.7/f$	45	$2.01+3.9/f$
4	$1.32+0.6/f$	14	$1.72+2.0/f$	24	$1.86+2.8/f$	50	$2.04+4.1/f$
5	$1.40+0.8/f$	15	$1.74+2.1/f$	25	$1.87+2.8/f$	60	$2.08+4.4/f$
6	$1.46+1.0/f$	16	$1.75+2.2/f$	26	$1.88+2.9/f$	80	$2.15+5.0/f$
7	$1.51+1.2/f$	17	$1.77+2.3/f$	27	$1.89+3.0/f$	100	$2.19+5.4/f$
8	$1.55+1.4/f$	18	$1.87+2.4/f$	28	$1.90+3.1/f$	200	$2.37+6.9/f$
9	$1.59+1.5/f$	19	$1.80+2.4/f$	29	$1.91+3.1/f$	500	$2.61+9.1/f$
10	$1.62+1.6/f$	20	$1.81+2.5/f$	30	$1.91+3.2/f$	∞	$\infty + \infty/f$

Many times independence of selection has to be attained by a more or less reasonable fiction, most often a fiction of a hypothetical population, (compare Section B4), to which a statistical step of inference may reasonably be sought. But if we are to have confidence in something beyond the bare limits, in space and in time, of what was observed, we must take nonstatistical steps of inference which are at least equally hard to support.

The typical character of a body of data on the basis of which we wish to establish statistical confidence is then a body of data divided in subbodies presupposed to show independence of fluctuation. Perhaps an intensive battery of psychological tests have been given to 500 children, and the result of a complex but specific calculation obtained. If it is reasonable to divide the 500 into ten 50's in a specified way, and then presuppose independence of fluctuation of the ten contributions corresponding to these ten groups, we can seek confidence in the overall complex result on the basis of the extent of quantitative agreement of the results for the 10 separate groups.

The observational situation may be such that we dare make up these groups at random. If this were so, and if the complex calculation were very simple, we would probably be able to "take down from the shelf" a conventional technique for confidence or significance. But it is much more frequent that, if the presupposition of independence of

Table 24

Example of use of multiple comparisons technique.

Enforced rest (minutes)	Average intercopulatory interval (seconds)
5	34.7
15	19.0
60	18.1
180	29.3

Estimated variance per rat = 68.3 $(\text{sec})^2$,

Estimated variance for mean of 11 rats = 6.21 $(\text{sec})^2$, (both on 33 degrees of freedom).

t-test for deviation of 15 second rest from 5 second rest:

$$t = \frac{1.90 - 34.7}{\sqrt{12.42}} = 4.45$$

$$t_{5\%} = 2.036 \text{ (from standard tables)}$$

$$t_{5\%}\,(s\sqrt{2}) = 2.036\sqrt{12.42} = 7.2 \text{ (seconds)}.$$

Hence *if no other times of rest had been used*

(a) these two would be significantly different at 5% and,

(b) we could have 95% confidence that the actual difference between intercopulatory intervals was within ±7.2 seconds of that actually observed, and was consequently between −22.9 and −8.5 seconds.

Since *four* times of rest were used, the factor from Table 23, amounting to

$$1.32 + \frac{0.6}{33} = 1.34$$

must be used, thus requiring an allowance of

$$(1.34)\,(7.2 \text{ seconds}) = 9.6 \text{ seconds}.$$

Thus we may have 95% confidence that *all* the statements of the class exemplified by

(ICI for 15 min.) − (ICI for 30 min.) = 19.0 − 18.1 ± 9.6 seconds
= between 8.7 and +10.5 seconds,

(ICI for 180 min.) − (ICI for 15 min.) = 29.3 − 19.0 ± 9.6 seconds
= between +0.7 and 19.9 seconds

(ICI for 15 min.) − (ICI for 5 min.) = 19.0 − 34.7 ± 9.6 seconds
= between −25.3 and −6.1 seconds

are correct.

fluctuation is to be reasonable, we must assemble the groups much more systematically. Perhaps we may divide the subjects into groups in accordance with the dates on which they were tested, or according to the sizes of the high schools attended, or according to the state or residence. The more obviously separated the groups, the more likely to be reasonable is the presupposition of independence of fluctuation.

H5. THE JACKKNIFE

The result of a simple computation, such as finding a mean or a slope, based upon a body of data divided into sub-bodies of independent fluctuation, can be provided with confidence limits by a procedure tailor-made for the purpose. The results of a complex computation, even if applied to a similarly divided body of data, cannot be so treated, because no firm of statistical tailors will have produced an appropriate special procedure. If a psychologist has given a battery of tests to some subjects in such a way that each item on each test can be scored in two ways, A and B, if he has then calculated split-half reliability coefficients for each combination of test and scoring method, and has averaged these reliabilities for each of the two ways of scoring, how is he to judge the significance of the difference of the two average reliabilities, especially since almost everything is correlated, to an unknown degree, with almost everything else? An honest estimate of significance must go back to differences between persons, or between groups of persons, since only here is independence of fluctuation at all reasonable. (Even here it may require a hypothetical population to make it reasonable.)

The simple approach to assessment of significance in such a situation is to repeat the complex calculation for each sub-body of data separately, and to use the spread of the results as a basis for judging the uncertainty of the result calculated from the whole body of data. This approach has various difficulties:

(1) if the sub-bodies are too small, the complex calculation may be impossible, as when it is sought to use a single point to determine a line;

(2) if the sub-bodies are somewhat larger, the complex calculation, though possible, may lead to results which:

   (a) are nonsensical; or

   (b) which vary too widely to be a fair basis for estimating the variability of the result of the complex calculation applied to the whole body of data.

In most circumstances, moreover:

(3) the results of a complex calculation will usually be biased, the amount of bias depending upon the size of the body of data used; the result for the whole data will usually not be free of bias.

It is worth some trouble to avoid these difficulties. Fortunately it is very easy to greatly reduce some, and eliminate the rest, by a simple device.

Suppose that there are r different sub-bodies, and that we are prepared to treat them as of equal weight. Let

$y_{(j)}$ = the result of applying the complex calculation to the whole body of data *with the exception* of the jth sub-body.

Let

y = the result of applying the complex calculation to the whole body of data, without exception.

Now define *pseudo-values* by

$$y_j' = ry - (r-1)y_{(j)} .$$

The price of carrying on to this point is no more than for the first suggestion. The complex calculation has to be gone through $r + 1$ times. (Especially with the rise of the electronic computer, the price in effort of such repetition goes down steadily.)

In most instances these pseudo-values, $y_1', y_2', \ldots, y_r'$ can now be treated, for the purpose of setting confidence limits, as if they were r individual observations on the result of the complex calculation, observations with independent fluctuations. This statement is far from obvious, but can be obtained and documented for a wide variety of instances by appropriate algebraic manipulation or by mathematical experimentation (Tukey 1957, Tukey and Chanmugan NYC2). Be the statement obvious or unobvious, it is surely useful, for there are many standard procedures for setting confidence limits for a population mean on the basis of a simple random sample of observations. Student's t is a classic, while such standard nonparametric procedures as the one-sample Wilcoxon test and the sign test are easily converted into confidence procedures.

H6. THE FEW EXCEPTIONS

All that remains is to list the cautions which need to be observed in the use of this all-purpose, Boy-Scout-jackknife-like confidence procedure. So far as is now known, there are only two broad classes of situations in which the jackknife may not be effective:

(1) situations in which the answer is coarse-grained; and

(2) situations in which the estimation is very narrow.

Both deserve a word of explanation.

If the result of the complex calculation behaved like "the most common number of children in a U.S. family" or, even more extreme, like "the number of U.S. presidents to be elected by the Democrats in 1964," where only a few values are at all possible (in the latter instance only two, i.e. "0" or "1"), this coarse-grainedness of answer is very likely to make the jackknife procedure ineffective.

If some one observation or some few observations dominate the value of the result, the result is said to be ***narrowly estimated.*** The largest observed value of some quantity is usually narrowly estimated, since this largest value is usually taken on in only one or a few of the instances observed. The average sales per outlet of a household gimmick sold only by Macy's, by Sears Roebuck, and by not very many small country stores is narrowly estimated, because it will be dominated by two observations, "How many does Macy's sell?" and "How many does Sears Roebuck sell?" There can be many other sorts of narrow estimation, but these two should identify the problem.

If the result for which confidence or significance statements are desired is narrowly estimated, the jackknife method is not likely to be effective. This is true whether the few dominant values are included among the actual observations or are only among the potential observations. In this latter case we speak of ***vanishing estimation.***

Thus if there is exactly one very extreme individual in a population, the fact that the most extreme observed value in the sample was represented by many individuals offers no protection. The sample offers little or no evidence as to how extreme the single anomalous individual may be.

Similarly, if a ***sample*** of sales for the household gimmick includes only sales for some country stores, the very much larger sales of Macy's and Sears Roebuck cause even more difficulty when they are missing than they would if they were in the sample.

When reasonable and possible, narrow or vanishing estimation is best avoided. If the very extreme value can, without loss of relevance and usefulness, be replaced by the value exceeded by only 1 instance in 100 much will have been gained. Not only will the jackknife method be applicable as soon as a sample of some hundreds is available, though a simpler special-purpose method should probably replace it, but the graspability of the result will be better in samples of any size. Rational study of the sales of the kitchen gimmick calls out for stratification into three strata of outlets: large department stores, large mail-order houses, small country stores. Once the data is thought of, and collected, in this way, there not only remains no problem of narrow estimation, but the whole process becomes much more efficient, helpful, and manageable. And so on.

It may not be possible to avoid narrow estimation, particularly when the analogs of Macy's or Sears Roebuck are unrecognized or unsuspected. If narrow estimation has to be faced, it must be regarded as a very special and very important difficulty, to be thought over carefully in each special instance.

EPILOGUE

Starting from the first badmandment, and its expansion stage by stage, we have discussed topics which may seem slightly disconnected, which now ought to be drawn together to a point, to be focused like the light rays of well-behaved instances of geometrical optics. At the same time, certain of the points made above can receive the additional emphasis they deserve.

Let us begin by trying to epitomize various sections of the discussion in terms of conclusions (and comments thereupon):

A1. Causation can only be established as a ***theoretically*** inevitable consequence of ***empirical*** observations. (Failure to recognize this dual requirement leads to asking too much of statistics, and to consequent dissatisfaction.)

A2. Those who regard the very arbitrary act of "doing nothing about · · · " as "not arbitrary" are afraid of being ***called*** arbitrary rather than of ***being*** arbitrary. (Thereby they lose many opportunities and are often very arbitrary.)

B1. From empirical observations to desired conclusion is two steps, and at most the first can be purely statistical. (It is best that the first step reach out as far as possible.)

B2. *Drosophila* often stand for all flies, all insects, or all life. (This must be a biological act of faith, not just a statistical one.)

B3. To regard the particular redheads, brunettes and blondes who entered a beauty contest as more than just a sample is unwise. (If they are not random samples from specifiable populations, that fact does not warrant promoting each subset of contestants to be a population.)

B4. Even a sociologist's single interconnected group of people needs to be considered a sample — often a sample of size one. (When so regarded, more efficient sampling designs become clear.)

C1. Noting and utilizing empirical regularities was very important in the growth of physics. (It may be expected to be just as important elsewhere.)

C2. Formal statistics, at least in science, exists as a relatively precise mode of communication, both inter- and intra-individual. (It is not just a way to make safe statements or a way to choose good bets.)

C3. While setting up an analysis in advance should be regular practice, the data must always have a chance to guide its own analysis. (Certain formal statistical difficulties can and should be overcome. But the investigator should not wait for this to happen before looking to the data for guidance.)

D1. Broad classes are not as useful as narrow ones. (Especially when later analysis is wisely performed and interpreted.)

D2. Classes so fine that two experts cannot agree on the exact class for three-quarters of the cases need not necessarily be too fine. (Enrico Fermi once said: "Measurement is just the making of fine distinctions".)

D3. "Controlling" variables in *broad* classes is not at all certain to be effective. (Of course it does help.)

D4. Dichotomizing instead of choosing the best scale one can select is most unwise, and brings a fool's reward. (An "arbitrary" equal-step scale is always better than a randomly chosen dichotomy.)

E1. Deep and careful searching into physical measurement has led to a formalization of what measurement in the highest monastic sense should be. (And this formalization is subject to misinterpretation!)

E2. Choosing the scale of the response without regard to the effects of the variables to be studied is almost sure to lead to trouble. (Massive "interactions" are the most likely trouble sign.)

E3. If lack of knowledge forced us to start with an arbitrary response scale, the first way to use the data is to seek for a better mode of expressing the response. (With luck such an empirical step can lead to deep and broad advances, but it will anyway be valuable on its own account.) If a scale of response can be found so that the factors act additively, the result will be joint measurement of the factors according to the highest monastic standards.

E4. Expressing relative numbers as a response is usually best done in terms which differ from percentages by stretching the "tails" (in comparison to the "center"); three modes of expression are often used for this purpose. (Graph papers with percentage scales makes the use of these three simple. Shape-changing of scales matters to later analysis, while uniform stretching or shrinking, as produced by a linear transformation, ordinarily does not.)

E5. These modes have rather understandable properties. (The use of the modes can be quite helpful even to those who do *not* know either these properties or their formal definitions.)

E6. In very simple examples, the use of these modes often exposes insights which percentages concealed. (Don't be blinded by the simplicity of such changes. The largest dividends are likely to come from simple ways of doing things better.)

F1. Borrowing strength from parallel, formally irrelevant observations is not only often practiced, but desirable. (Its dangers are better faced than avoided.)

F2. The technique of pooling information "within" portions of a classification (as opposed to pooling the raw counts themselves) is an important way of borrowing strength. (It is much used, in concealed form, in modern analysis of measurements. It offers many possibilities of better analysis in behavioral science.)

F3. Statistically appropriate "adjustment" for the values of variables whose effects are confusing and irrelevant is a valuable tool. (It is, in large part, another way of borrowing strength. Its use must, of course, be reported.)

F4. Residuals, which represent differences between what was observed and what has been systematically described, offer the greatest possibility of discovering unexpected things in a body

of data. (In a concealed way, they underlie many statistical procedures.)

F5. Counted data are subject to adjustment, and can generate residuals, once attention is given to subgroups, instead of individuals. (Any of various modes of expression for the relative numbers showing a characteristic in each subgroup may prove most helpful.)

F6. Data quite frequently, perhaps usually, has to be analyzed in stages if its analysis is to be either efficient or searching. (Often pre-analysis of small subsamples is desirable. Sometimes a "complete" analysis provides only a jumping-off place for a better analysis.)

G1. Many fallacies, which should be carefully studied and avoided, are prevalent in the use of tests of significance. (But even so, the overall value of such tests is great.)

G2. There is much to be gained by avoiding "omnibus" tests through dissection of intercomparisons into more meaningful pieces. (And by treating "error rate" as a scarce commodity, one to be carefully and wisely allocated.)

G3. In many specific instances, splitting a chi-square offers very much more insight into what is happening. (As well as often converting a lack of significance into significance.)

G4. There are now adequate procedures for making all possible intercomparisons among a set of means or other typical values. (These are statistically respectable, and involve meaningful probability statements.)

T3. More attention needs to be given to the multiple uses to which tabulations will be put. (And a suggestion by Dwyer offers one approach.)

Most of these epitomes can be classified under one or more (average 1.7) of the following prime *good*mandments (classification indicated):

(X) *As a scientist and investigator you can never give over your responsibilities as a thinking, judging, noticing, feeling person.* You can receive much help from such tools as concepts and statistical techniques, but you must use them, not let them use you. You can do better than a machine, but *only* by taking the chance of doing worse. (A1, A2, B1, B2, B3, B4, C1, C2, C3, E1, E4, E5, F1, F6, G1, G2, G3.)

(Y) *The twin arts of empirical approximation and statistical inference complement each other.* Either alone yields limited gains and exposure to certain dangers. Both together offer far greater returns (and less danger). (A2, C1, C3, D3, D4, E2, E3, E4, E5, E6, F1, F2, F3, F4, F5, F6, G2, G3, G4.)

(Z) *You must "sit loose" to data, to results of analyzing data, and to interpretations of these results, if you are to get full value from any of them.* Treating any one of these as "black or white" means discarding both information and opportunities for insight. (A1, B1, B2, B3, B4, C2, D1, D2, E2, F3, F4, F6, G1, G2, G3, T3.)

Of what are these three prime goodmandments extensions? To what focus can we finally converge? There is but one choice:

IN BUILDING NEW SCIENCES, LOOK TO HOW THE ELDER SCIENCES *ACTUALLY* WERE BUILT.

Do not look to how it is stated that they should have been built, or to how their completed edifices appear, even though, as is often the case, one or other of these is claimed to show "how they were built". Look to the formative stages of the elder sciences, look to the actual practice of scientists during those stages.

R. REFERENCES AND BACKGROUND MATERIAL

The sections which follow are an attempt to try to direct the interested reader to material which extends, or illuminates, or contradicts the material of the previous sections. The last section (R9) gives details of all references cited. Background material for the appendices (S, T, U, V, W) is to be found at the end of each appendix.

R1. BACKGROUND FOR CHAPTER A

There seem to be few discussions, if any, of the *establishment* of causal relations. The classical technique, still not widely enough employed, for the utilization of assumed causal relations in untangling complex numerical data is Sewell Wright's path analysis. Wright 1923 and Wright 1921 are good introductions. Wright 1934 and Wright 1951 give advanced accounts. Path analysis was originally stated in correlation form, but can be put into regression terms (Tukey 1954). For recent expositions see Li 1955, Li 1956, Turner and Stevens 1959.

At the time of its introduction, path analysis was attacked (Niles 1922, 1923) by proponents of Karl Pearson's (1892 ff) view (which has a long philosophical history) that causation was merely close correlation. (Fortunately this attempt to find certainty in uncertainty seems to have lost its popularity among users of quantitative method, though it seems still to be popular among philosophers.)

For a discussion of the meaning of causation, see Wold 1966. (The writer would agree with Wold about meaning, but not about ease of verification.) Questions of antecedent and intervening variables are often relevant in connection with causality. Lazarsfeld 1958 offers an introduction to their use (at pp. 117-124 and 130).

Even less can be offered as background on what does and does not constitute arbitrariness.

R2. BACKGROUND FOR CHAPTER B

The opposite view to that put forward in this chapter has been strongly stated by Kempthorne (1955, 1961).

At a more technical level, discussions of the "corrected error term" in analyses of variance are related to this question. See Chapter 5 of Goulden 1952 or Section 11.8 of Snedecor 1946 for discussions in an agricultural context, Green and Tukey 1960 for a partial discussion in a psychological context, and Fisher 19?? (reference lost!) for an early and fundamental statement. (Still more technical material may be found in Wilk and Kempthorne 1955 and 1956, and in Cornfield and Tukey 1956.)

R3. BACKGROUND FOR CHAPTER C

Again a dearth of references.

R4. BACKGROUND FOR CHAPTER D

Surely there must be many references in the psychological literature. But where?

The effects of grouping on simple statistics is a classical topic. The effects of grouping on normal variates are among the topics treated by Fisher in "On The Mathematical Foundations of Theoretical Statistics" (see Fisher 1922, pp. 317-321). The connection between grouping efficiency and reclassification agreement seem to have been first discussed in Tukey 1950.

The subject of selecting effective scales for ordered classifications, with or without additional information, has been undertaken by Abelson and Tukey (1963, others in preparation).

The limitation of statistical procedures by scale type has been almost exclusively discussed by Stevens (1946, 1951, 1955, 1959). See also Mosteller (1958).

R5. BACKGROUND FOR CHAPTER E

Discussions of fundamental measurement seem to have been largely confined to Campbell 1920, 1928, and Stevens 1946, 1951, 1959. It is clearly time for a reconsideration of the whole subject. (A brief discussion will be included in Tukey NYC1.)

The choice of modes of expression, and the reasons for choice, have also been rather neglected. A fair amount of general discussion together with rather careful consideration of techniques for choosing an appropriate mode may, in due course, appear in Tukey NYC1. The earlier literature speaks mainly of "transformation" (e.g., Bartlett 1947).

For further discussion of particular modes of expressing counted fractions, see Appendix U below.

For a brief discussion of modes of expression for other quantities than counted fractions, see Appendix V below.

R6. BACKGROUND FOR CHAPTER F

Explicit discussions of "borrowing strength" and "pooling within" seem notable by their absence, although these notions underlie most of the refined procedures of modern statistical analysis.

Similar remarks seem to apply to the other topics treated in Chapter F.

The use of *sophisticated* statistical techniques, rather than simple ones illustrated in Chapter F is *not* likely to be often necessary in connection with "pooling within" counted fractions expressed in normits, logits, etc. Moderately extensive examples of sophisticated techniques may be found in Yates 1955.

R7. BACKGROUND FOR CHAPTER G

Many of the main references on when significance tests should be used were already cited in Section G1. The discussion begun by Selvin 1957 was continued by McGinnis 1958 and Kish 1959 (as well as by

shorter discussions mentioned under Selvin 1957). Wold 1956 approaches the question somewhat differently. The discussions in Merton et al. 1958 (at pages 302-304) and Zeisel 1955 were cited above. A strong opposing view has been taken in Kempthorne 1961.

The interpretation of the growth of statistics as a search for certainty does not seem to appear in the literature.

Many books on laboratory technique discuss the importance of systematic errors, but the writer knows of none that goes on to discuss their implications for the analysis of fluctuating errors in any detail. Some useful general background may come from reading Wilson 1952 on general scientific technique, and DuMond and Cohen 1958 on the present state of knowledge of the fundamental physical constants. (Their statement at page 7-164 that "the adjusted 'best' values have changed in the last two years by from five to nine times the estimated probable errors of the December 1950 evaluation" clearly illustrates the importance of systematic errors in physical science measurements.)

Many books on statistics devote some space to the fallacies of significance testing. None can be complete, since new fallacies are frequently invented to supplement the old. It seems best not to suggest any particular sources.

R8. BACKGROUND FOR CHAPTER H

The splitting and allocation of error rates has been little discussed in print, perhaps because of its necessary use of judgment, and its consequent apparent arbitrariness. The best discussion of the partition of chi-square is undoubtedly that of Cochran (1954) which is *not* made obsolete by more detailed work of Lancaster.

Multiple comparison procedures are a recent development, and have led to strongly conflicting views. The only exposition of views similar to the writer's is to be found in Ryan 1959a, which is addressed to psychologists. (It is hoped the Kurtz et al. 1965 will appear shortly and that Tukey 1960u will appear.) Extended lists of references to articles presenting various views can be found in Kurtz et al. 1965.

The only direct reference to the jackknife procedure in print is Tukey 1958. A relatively full account is being prepared by Chanmugam and Tukey NYC2.

R9. REFERENCES CITED

Abelson, Robert P. and Tukey, John W. (1963). Efficient utilization of non-numerical information in quantitative analysis: General theory and the case of simple order, *Annals of Math. Statistics* 34, 1347-1369.

Bartlett, M. S. (1947). The use of transformations, *Biometrics* 3, 39-52.

Box, G. E. P. (1957). Abstract #407. Iterative experimentation, *Biometrics* 13, 240-241.

Campbell, Norman Robert (1920). *Physics: The Elements (Vol. 1).* Cambridge University Press. (Reissued in 1957 as *Foundations of Science: The Philosophy of Theory and Experiment.* Dover, New York.)

Campbell, Norman Robert (1928). (An account of the principles of) *Measurement and Calculation.* Longmans, Green: London.

Carington, (Walter) Whately (originally Walter Whately Smith) (1945). *Telepathy; an Outline of its Fact, Theory and Implications.* Methuen, London.

Chung, J. H. and Fraser, D. A. S. (1958). Randomization tests for multivariate two-sample problem, *J. Amer. Statist. Assoc.* 53, 729-735.

Cochran, W. G. (1954). Some methods for strengthening the common chi-square tests, *Biometrics* 10, 417-451.

Cornfield, Jerome and Tukey, John W. (1956). Average values of mean squares in factorials, *Annals of Math. Statistics* 27, 907-949.

Dempster, A. P. (1960). A significance test for the separation of two highly multivariate small samples, *Biometrics* 16, 41-50.

DuMond, Jesse W. M. and Cohen, E. Richard (1958). Fundamental Constants of Atomic Physics. *Handbook of Physics.* (E. U. Condon and H. Odishaw, eds.) Part 7, Chapter 10, 7-143 to 7-173. McGraw-Hill, New York.

Dwyer, Paul S. (1942). Grouping methods, *Annals of Math. Statistics* 13, 135-155.

Fisher, R. A. (Sir Ronald) (1922). On the mathematical foundations of theoretical statistics, *Phil. Trans. Roy. Soc. London* A222, 309-368. (Reprinted as paper 10 in: Fisher, R. A. (1950). *Contributions to Mathematical Statistics.* Wiley, New York.

Fisher, R. A. (Sir Ronald) (1935ff). *The Design of Experiments.* Edinburgh, Oliver and Boyd. (Hafner, New York) 6th edition, 1951.

Friedman, Milton (1957). *Theory of Consumption Function.* Princeton University Press.

Gaito, John (1959). Multiple comparisons in analysis of variance, *Psychol. Bull.* 56, 392-393.

Glueck, Sheldon and Glueck, Eleanor (1950). *Unravelling Juvenile Delinquency.* Commonwealth Fund, New York.

Goulden, Cyril H. (1952). *Methods of Statistical Analysis.* 2nd edition. Wiley, New York. (Chapman and Hall, London).

Green, Bert F., Jr. and Tukey, John W. (1960). Complex analyses of variance: general problems, *Psychometrika* 25, 127-151.

Hammersley, John M. (1954). Poor man's Monte Carlo, *J. Roy. Statistical Soc.* B16, 23-26 (discuss. 61-75).

Hempel, Carl G. (1952). Fundamentals of concept formation in empirical science. *Intern. Encycl. of Unified Science.* Vol. 2, Part 7. Chicago University Press.

Hyman, Herbert H. (1955). *Survey Design and Analysis.* Free Press, New York.

Johnson, Palmer O. (1949). *Statistical Methods in Research.* (especially 298-310). Prentice-Hall, New York.

Johnson, Palmer O. and Tsao, Fei (1945). Factorial design in the determination of differential limen values, *Psychometrika* 9, 107-144.

Kempthorne, Oscar (1955). The randomization theory of experimental inference, *J. Amer. Statist. Assoc.* 50, 946-967.

Kempthorne, Oscar (1961). The design and analysis of experiments with some reference to educational research. *Research Design and Analysis, Second Annual Phi Delta Kappa Symposium on Educational Research.* 97-126. Phi Delta Kappa, Inc., Ames, IA.

Kish, Leslie (1957). Confidence intervals for clustered samples, *Am. Sociol. Rev.* 22, 154-165.

Kish, Leslie (1959). Some statistical problems in research design, *Am. Sociol. Rev.* 24, 328-338.

Kurtz, Thomas E., Link, Richard, F., Tukey, John W. and Wallace, David L. (1965a). Short-cut multiple comparisons for balanced single and double classifications: Part I, Results, *Technometrics* 7, 95-161.

Kurtz, Thomas E., Link, Richard F., Tukey, John W. and Wallace, David L. (1965b). Short-cut multiple comparisons for balanced single and double classifications: Part II, Derivations and approximations, *Biometrika* 52, 485-498.

Landes, David S. (1954). Social attitudes, entrepreneurship, and economic development: A comment, *Explorations in Entrepreneurial History* 6, 245-272. (esp. Appendix Table 2)

Lazarsfeld, Paul F., Berelson, Bernard and Gaudet, M. (1948). *The People's Choice.* 2nd Edition. Columbia University Press.

Lazarsfeld, Paul F. (1958). Evidence and inference in social research, *Daedalus* 87, 99-130.

Leverett, Hollis M. (1947). Table of mean deviates for various portions of the unit normal distribution, *Psychometrika* 12, 141-152.

Li, Ching Chun (1955). *Population Genetics.* University of Chicago Press.

Li, Ching Chun (1956). The concept of path coefficient and its impact on population genetics, *Biometrics* 12, 190-209.

Luce, R. Duncan (1959). On the possible psychophysical laws, *Psychological Review* 66, 81-95.

McCall, W. T. (1939). *Measurement.* MacMillan, New York.

McGinnis, Robert (1958). Randomization and inference in sociological research, *Am. Sociol. Rev.* 23, 408-414.

Merton, Robert K., Reader, George G. and Kendall, Patricia L., eds. (1957). *The Student-Physician.* Harvard University Press.

Morgenstern, Oscar (1950). *On the Accuracy of Economic Observations.* Princeton University Press.

Mosteller, F. (1958). The mystery of the missing corpus, *Psychometrika* 23, 279-289.

Niles, H. E. (1922). Correlation, causation, and Wright's theory of "path coefficients", *Genetics* 7, 258-273.

Niles, H. E. (1923). The method of path coefficients: an answer to Wright, *Genetics* 8, 256-260.

Pearson, E. S. (1939). William Sealy Gosset, 1876-1937. "Student" as a statistician, *Biometrika* 30, 210-250.

Pearson, Karl (1937). *The Grammar of Science.* J. M. Dent and Sons, London.

Popham, Robert E. (1953). A critique of the genetotrophic theory of the etiology of alcoholism, *Q. J. Studies on Alcohol.* 14, 228-237.

Reichenbach, Hans (1951). *The Rise of Scientific Philosophy.* University of California Press, Berkeley.

Ryan, T. A. (1959a). Multiple comparisons in psychological research, *Psychol. Bull.* 56, 26-47.

Ryan, T. A. (1959b). Comments on orthogonal components, *Psychol. Bull.* 56, 394-396.

Selvin, Hanan J. (1957). A critique of tests of significance in survey research, *Amer. Sociol. Rev.* 22, 519-527. (See also in *Amer. Sociol. Rev.* 23, 85-86 (Gold), 86 (Selvin), 199 (Beshers), 199 (Selvin), 408-414 (McGinnis).)

Snedecor, George W. (1946). *Statistical Methods.* 4th Edition. Iowa State College Press, Ames, IA.

Stevens, S. S. (1946). On the theory of scales of measurement, *Science* 103, 677-680.

Stevens, S. S. (1951). Mathematics, measurement and psychophysics. *Handbook of Experimental Psychology.* (S. S. Stevens, ed.) 1-49. Wiley, New York.

Stevens, S. S. (1955). On the averaging of data, *Science* 121, 111-116.

Stevens, S. S. (1959). Measurement, psychophysics, and utility. *Measurement: Definitions and Theories.* (C. W. Churchman and P. Ratoosh, eds.) Chapter 2, 18-63. Wiley, New York.

Stouffer, Samuel A., Guttman, Louis, Suchman, Edward A., Lazarsfeld, Paul F., Star, Shirley A. and Clausen, John A. (1950). *Measurement and Prediction,* Volume IV of Studies in Social Psychology in World War II. Princeton University Press.

Student (1938). Comparisons between balanced and random arrangements of field plots, *Biometrika* 29, 363-379.

Tukey, John W. (1950). Discussion, *J. Clinical Psychol.* 6, 61-74.

Tukey, John W. (1954). Causation, regression and path analysis. *Statistics and Mathematics in Biology.* (O. Kempthorne, T. A. Bancroft, J. W. Gowen, and J. L. Lush, eds.) Chapter 3, 35-66. Iowa State College Press, Ames, IA.

Tukey, John W. (1954). Comparing two small samples on many items. *Memorandum Report 54,* Statistical Research Group, Princeton University.

Tukey, John W. (1957). On the comparative anatomy of transformations, *Annals of Math. Statistics* 28, 602-632.

Tukey, John W. (1958). Bias and confidence in not quite large samples (abstract), *Annals of Math. Statistics* 29, 614.

Tukey, John W. (1960). A survey of sampling from contaminated distributions. *Contributions to Probability and Statistics.* (Essays in honor of Harold Hotelling) (I. Olkin, S. G. Ghurye, W. Hoeffding, W. Madow and H. B. Mann, eds.) 300-327. Stanford University Press.

Tukey, John W. (NYC1). Choice and change of modes of expression. Not yet (1985) completed.

Tukey, John W. (1960u). The problem of multiple comparisons. In preparation. (Dittoed version circulated for comment in 1953.)

Tukey, John W. and Chanmugam, J. (NYC2). Approximate confidence estimates for most estimates. Not yet (1985) completed.

Turner, M. E. and Stevens, C. O. (1959). The regression analysis of causal paths, *Biometrics* 15, 236-258.

Wilk, Martin B. and Kempthorne, Oscar (1955). Fixed, mixed and random models, *J. Amer. Statist. Assoc.* 50, 1144-1167.

Wilk, Martin, B. and Kempthorne, Oscar (1956). Some aspects of the analysis of factorial experiments in a completely randomized design, *Annals of Math. Statistics* 27, 950-985.

Williams, Roger J., Beerstecher, Ernest Jr., Sutton, H. Eldon, Berry, Helen Kirby, Brown, William Duane, Reed, Janet, Rich, Gene B., Berry, L. Joe and Williams, Roger (1950). Biochemical individuality V. Exploration with respect to the metabolic patterns of compulsive drinkers, *Archives of Biochemistry* 29, 27-40.

Wilson, E. Bright, Jr. (1952). *An Introduction to Scientific Research.* McGraw-Hill, New York.

Wold, Herman O. A. (1956). Causal inference from observational data, *J. Roy. Statist. Soc.* A119, 28-61.

Wold, Herman O. A. (1961). Unbiased predictors. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* 31, 719-761.

Wold, Herman O. A. (1966). On the definition and meaning of causal concepts. *Model Building in the Human Sciences. Entretiens de Monaco en Sciences Humains, session 1964,* (R. Peltier and H. Wold, eds.) 265-295. Centre International d'Etude des Problémes Humaine, Monaco.

Working, Holbrook (1933). Price relations between July and September wheat futures at Chicago since 1885, *Wheat Studies* 9, 187-238.

Working, Holbrook (1934). Price relations between May and newcrop wheat futures at Chicago since 1885, *Wheat Studies* 10, 183-230.

Wright, Sewall (1921). Correlation and causation, *J. Agric. Res.* 20, 557-585.

Wright, Sewall (1923). The theory of path coefficients: A reply to Nile's criticism, *Genetics* 8, 239-255.

Wright, Sewall (1934). The method of path coefficients, *Annals of Math. Statistics* 5, 161-215.

Wright, Sewall (1951). The genetical structure of populations, *Annal of Eugenics* 15, 323-354.

Yates, Frank (1951). The influence of *Statistical Methods for Research Workers* on the development of the science of statistics, *J. Amer. Statist. Assoc.* 46, 19-34.

Yates, Frank (1955). The use of transformations and maximum likelihood in the analysis of quantal experiments involving two treatments, *Biometrika* 42, 382-403.

Zeisel, Hans (1955). The significance of insignificant differences, *Public Opinion Quarterly* 19, 319-321.

S. THE ARITHMETIC OF GROUPING

Rounding off numerical values, grouping frequency distributions, and classifying on the basis of either rules or judgment, lead to essentially similar problems of how much information is lost by severe rounding, long grouping intervals, or broad classifications, and of how much effort is wasted by overgentle rounding, very short grouping intervals, or overnarrow classifications. The best extent to round, group, or classify has to be learned from essentially similar facts. Yet in much of today's practice we are too "cautious" in all three. And by being "cautious" we adopt wholly inconsistent standards, frequently carrying *more* decimal places than we need, and very often refusing to make as *fine* classifications as would help us. The first of these three practices tends to deviate in one sense from our relatively sound practice in grouping frequency distributions, the third tends to deviate oppositely.

S1. KINDS OF ROUNDING, AND SOME PROPERTIES

The question of rounding off values is always with us in one form or another. It is often helpful to know just what the quantitative effects of rounding are. Suppose we replace continuously, and sensibly uniformly, distributed values by values rounded to steps of width h. (Rounding to 3 decimal places, for example, corresponds to $h = 0.001$.) How much shift, on the average, is there between unrounded value and rounded value?

Kind of Rounding	Average Square of Shift
Perfect (to *nearest* rounded value)	$(1/12)h^2$
Random (to *either* of nearest two rounded values with equal probability)	$(4/12)h^2 = (1/3)h^2$
Perverse (to *further* of nearest two rounded values)	$(7/12)h^2$

At least three answers are helpful, viz: These three answers are worth considering because they are easy to think about, and because they bracket most usual situations which arise either when rounding given numbers, or when measuring "to the nearest · · · ."

How large can h^2 reasonably be? And why? The why must usually come from some comparison of average square of shift due to rounding with the average square of the fluctuations arising from other sources.

Many of the books on elementary statistics which discuss the calculation of means, variances, etc., of *large* bodies of data through the formation of a grouped frequency distribution indicate the desired fineness of grouping (here corresponding to perfect rounding) by suggesting that 10 to 20 cells of the frequency distribution should be occupied. If the distribution is crudely normal, the range of occupied cells will cover some 5 or 6 standard deviations. Let us take 5σ as a convenient number. If this is $20h$ or $10h$, then $h = \sigma/2$ or $\sigma/4$, and $(1/12)h^2$ is $\sigma^2/48$ or $\sigma^2/192$, corresponding to an increase of mean square fluctuation due to grouping of 2% of σ^2 or 0.5% of σ^2. When we allow a little for longer occupied ranges due to the non-normality of many practical distributions, it seems right to judge that tacit statistical

practice through the years has found 5% to 1% increase in mean square fluctuation due to grouping entirely palatable. This is not different from what we might have expected if we had approached the question without background of experience.

S2. *ROUNDING NORMAL DISTRIBUTIONS*

In one of his early long and path-breaking papers, R. A. Fisher (1922) studied the effect of (perfect) rounding on (perfectly) normal distributions. His results were surprising, and yet have been typically overlooked. In addition to the average effects of rounding known as Shepard's corrections which do not depend upon how much the population mean must be rounded to reach a round value, he found small effects depending on the relation of the population mean to the two nearest round values. In samples of less than 12,000 million million this effect is less than 1/10th the standard deviation of the sample mean. (Similar results hold for all four of the first four moments in samples of less than a million million.)

These results cannot be taken over directly for practical guidance, since rounding has a somewhat greater effect upon many practical distributions than it has upon perfectly normal distributions. But they serve as an excellent remedy for the feeling that grouping or rounding is always dangerous.

S3. *HOW MANY DECIMAL PLACES DO EXPRESSIONS OF COUNTED FRACTIONS REQUIRE?*

Table 10 in Section E4 presented values of anglits, (modified) normits, and (modified) logits to only two decimal places. Table 11 presented values of "numerator" where simple random sampling variance $\doteq \frac{\text{numerator}}{\text{sample size}}$ for these same modes of expression. It is clear from the table that, for these three modes, "numerator" is never less than unity, so that the simple random sampling variance is never less than 1/(sample size).

As we have seen, if the distribution of some quantity is relatively smooth, rounding to steps of width h introduces rounding of variance about $h^2/12$. With $h = .01$, as in our case, this rounding variance is 0.0000083^+. This is a rather small variance.

The rounding variance will almost surely be accepted so long as it is no more than 5% of the random sampling variance. Thus the random sampling variance needs to be at least 0.00017 = 1/6000 if this much rounding variance is to be acceptable. Samples of size no more than 6000 will have this property.

If we require a rounding variance of no more than 1% of random sampling variance, a similar calculation shows that it is sufficient for the sample size to be no larger than 1200.

In practice, very large samples are rarely so conducted that the variability associated with simple random sampling dominates the variability of the answers. Accordingly, the precision offered by two decimal place accuracy in anglits or modified normits is usually quite sufficient for practical samples of any size, while that offered by half-logits is even more certain to be sufficient.

For the "doubled fraction," (column (1) in Table 10), the "numerator" becomes quite small for extreme fractions. For such fractions, two decimals in the doubled fraction may not suffice for samples of some hundreds.

S4. *WHEN DOES IT PAY TO SPLIT A BROAD CLASS INTO TWO NARROW ONES?*

In Section D2 we argued strongly that it was usually wise to split a broad classification if one had any reasonable basis for doing so. In this appendix we seek to provide concrete support for this position by treating some simple examples. These examples are not supposed to represent exactly what happens in any one actual instance of broad classification. They are supposed to provide specific instances which will help the reader in thinking about the broad class of problems. (After all, the methods we use are simple; each reader who cares to can choose his own examples and treat as many of them as he desires in a similar way.)

We have no hesitation in seeking a simple illustrative and informative situation by assuming several things, no one of which necessarily holds in practice, namely:

(1) There is an underlying true value for what is being classified;

(2) This true value can be wisely expressed in some definite quantitative terms;

(3) In these quantitative terms, the distribution of the true value is sensibly rectangular;

(4) The boundaries of the broad classes are infinitely sharp; this classification is perfect;

(5) If a broad class is divided into two narrow classes to each of which is assigned a score, or if the same score is assigned to each item in the broad class, the proper measure of unsatisfactoriness of classification is the mean square error, the average squared difference between score and true value.

Once these assumptions have been made, there is no loss of generality in assuming that the true values of the broad class extend from 0 to 1. If the natural choice of score for the undivided broad class, which here is the best, is made, namely 1/2, then the mean square error will be 1/12 = 4/48.

For perfect splitting, in which all items with true values between 0 and 1/2 fall into the lower class, while all items with true values between 1/2 and 1 fall in the upper class, the natural scores are 1/4 for the lower class and 3/4 for the upper. The mean square error will be $(1/2)^2(1/12) = 1/48$ for each class separately, and hence the same for the combination.

For random splitting between upper and lower halves, the mean square errors around 1/2 will be 1/12 for each class, and each class will have 1/2 for its mean. If we allow "each class to find its own level," so that each is scored 1/2, the overall mean square error remains 1/12 = 4/48. If we force "equally spaced" scores of 1/4 and 3/4 upon the halves, as is not unlikely, this figure must be increased by $\frac{1}{2}\left(\frac{1}{4}-\frac{1}{2}\right)^2+\frac{1}{2}\left(\frac{3}{4}-\frac{1}{2}\right)^2 = 1/16 = 3/48$, to reach 7/48.

Thus we have obtained most of the numbers in the following list of mean square errors

	Equally Spaced Scores (1/4 and 3/4)	2:1 Scores*	Finding Own Level
Perfect Split	1/48	1/36	1/48
No Split	4/48	4/48	4/48
Random Split	7/48	5/45	4/48

* Corresponding to scores of 1/3 for lower class and 2/3 for upper class.

We see that if we can let the upper and lower halves "find their own level" there is no loss from random splitting, while even if we rigidly impose equally spaced scores on the halves, the random split is only as much worse than no split, namely 7/48 − 4/48 = 3/48 = 4/48 − 1/48, as the perfect split is better. (With compromise scores at 1/3 and 2/3, even random splitting is not very expensive, while perfect splitting produces a very considerable gain.)

The notions of "perfect split," "no split," and "random split" are quite clear, but what we need to consider most are intermediate cases. When we do this, the notion of % classification discrepancy will help us somewhat. Consider any pair of items which have the same true value, and hence fall in the same broad class. When they are assigned to the narrow classes, they may both be assigned to some one half, or they may be assigned one to each half. The latter situation we call a classification discrepancy, and we ask what fraction of the pairs with identical true values will be discrepant. For our three leading instances the answer is easy.

0% classification discrepancy = for no split, and for perfect split;
50% classification discrepancy = for random split.

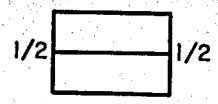
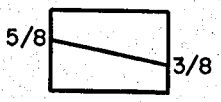
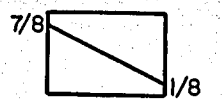
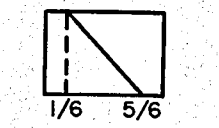
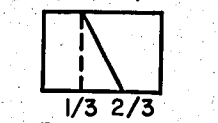
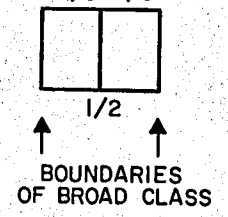
These are the extreme limits.

The behavior of intermediate cases is naturally described in terms of a splitting curve, which shows for each true value from 0 to 1 the chance that an item with such a true value will belong to each class. Table 25 shows % classification discrepancy, shape of splitting curve, and, for each of our three assumptions about scores, mean square errors, all for a variety of examples where the splitting curve is made up of straight lines.

As we saw above, if scores are allowed to find their own level we never lose by splitting. If, instead, we force scores of 1/3 and 2/3 on the halves, then one break-even situation arises when units with the highest or lowest true values (1 or 0) have one chance in four of being assigned to the wrong narrow class. And if, as an extreme, we force scores of 1/4 and 3/4 on the halves, one break-even situation arises when such extreme items have one chance in eight of being assigned to the wrong narrow class.

Table 25

Effect, when there is an Underlying Continuous (and Uniform) Distribution of "True Values," of Splitting One Broad Class into Two Narrow Classes with Varying Precision and with Varying Assignment of Scores

Fraction of Classification Discrepancies	Shape of Splitting Curve	Ratio of Mean Square Error of Classification After Subdivision to that before Subdivision: If treated as equally spaced	If treated as spaced 1:2	If allowed to find own spacing
50%		1.75	1.11	1.00
49%		1.50	1.06	0.97
46%		1.25	1.00	0.94
41%		1.00	0.94	0.92
33%		0.75	0.89	0.89
22%		0.47	0.36	0.32
11%		0.30	0.33	0.30
0%		0.25	0.33	0.25

When we recall that, in terms of our original example in Section D2, this break-even situation corresponds to situations where an individual who is truly on the *very lower edge of the middle class,* right next the upper working class, has, in fact, one chance in four, or one chance in eight, of being assigned to the upper middle class, we see just how poor our classifying ability must be if we are to break even, instead of gaining, when we split a broad class into two narrow ones.

Table 26

Chances of Shifts of Varying Numbers of Classes on Independent Reclassification into Classes of Varying Fineness

Classes per Standard Deviation of Judgment*	Chances of Shifts of 0	±1	±2	±3	more	Efficiency of Classification**
0.35	61%	38%	1%			60%
0.47	36%	60%	4%			73%
0.71	37%	48%	13%	2%		85%
1.4	29%	25%	24%	13%	8%	96%
2.4	12%	21%	23%	15%	30%	98.5%
3.5	8%	16%	15%	14%	47%	99.3%

* or perturbation

** as ratio of original variance to grouped variance (1 to $1 + \frac{1}{12}\left(\frac{1}{.35}\right)^2$ for the first time)

S5. RECLASSIFICATION AGREEMENT AND EFFICIENCY

The discussion of the last section was devoted to the splitting of an extremely clearly defined class. It clearly provides a basis for deciding whether or not to split the middle class. There is also a place for a basis for answering analogous questions in situations where no one of the possible boundaries is better defined or more precise than another, where classification resembles quantitative measurement.

Table 27

Formulas Underlying Table 26

With variance of perturbation of judgment $= \sigma^2$; interval length $= \tau\sigma$; bases for classification and reclassification $= \sigma x$ and $\sigma x + \sigma y$ with $y \geqslant 0$ and $t =$ fractional part of x/τ,
then

$$\text{distribution of } y \sim \sqrt{\frac{1}{\pi}}\, e^{-y^2/4}\, dy, \quad y \geqslant 0;$$

$$\text{distribution of } t \sim dt, \quad 0 \leqslant t \leqslant 1;$$

$$y \text{ and } t \text{ are independent,}$$

and

$$\text{number of intervals shifted} = \text{integer part of } t + y/\tau,$$

so that the probability of shifting $\leqslant J$ intervals is

$$\sqrt{\frac{1}{\pi}} \int_0^{J\tau} e^{-y^2/4}\, dy + \sqrt{\frac{1}{\pi}} \int_{J\tau}^{(J+1)\tau} \left[J + 1 - \frac{y}{\tau}\right] e^{-y^2/4}\, dy$$

$$= 2\,\text{Gau}((J+1)\tau/\sqrt{2}) - 1 + 2J\left[\text{Gau}((J+1)\tau/\sqrt{2}) - \text{Gau}(J\tau/\sqrt{2})\right]$$

$$- \frac{2}{\tau}\sqrt{\frac{1}{\pi}}\, e^{-(J\tau)^2/4}\left[1 - e^{-(2J+1)\tau^2/4}\right]$$

which for $\tau = 2.8571$ gives

$$\text{for } J = 0\text{: } .9566 + 0.0 - .3436 = .6130,$$

$$\text{for } J = 1\text{: } .9999 + .0433 - .0513 = .9919,$$

so that the probabilities of shift of 0, shift of ± 1 and shift of ± 2 are, when rounded, 61%, 38% and 1% as in the top line of Table 25 ($.35 = 1/2.8571$).

For this second basis it is reasonable to consider the conventional prototype situation: True values smoothly distributed along a continuous scale; apparent values obtained from true values by additive normally distributed perturbations; apparent values sliced up by equally-spaced cell boundaries; cell lengths narrow with respect to width of distribution; perturbations in reclassification independent of

those in classification. To finish specifying the situation, we need only specify one more parameter,

$$\frac{\text{standard deviation of perturbation}}{\text{cell length}}$$

If we adopt various values for this parameter, we obtain the results in Table 26. (These are based upon the formulas summarized in Table 27.)

We see that if no more than 15% of independent reclassifications neither check, *nor even fall in a class adjacent to the original class,* at least 15% efficiency is lost because the classes are so broad.

T. TRANSMISSION OF QUANTITATIVE INFORMATION

The purposes of technical discourse are not unified. Most technical writings are intended to span at least a modest portion of the broad spectrum from what can be read by the general public with ease to what the highly-trained specialist can only puzzle out slowly and painfully. Verbal expressions are used to transmit information at varying degrees of complexity and sophistication; numbers, tables, charts and graphs, all the forms of quantitative expression, must be expected to do the same.

Sometimes a quantitative expression should convey a very general message to a nonspecialized reader. Sometimes a quantitative expression should convey modest detail, or even considerable detail, in a form which may safely be handled by the naive specialist, even perhaps by a misguided one. At other times a quantitative expression should convey complete detail in a form which may only be safely handled by experts. It is as much a mistake to expect a single quantitative expression to meet all these requirements as it is to expect a single verbal expression to meet correspondingly diverse ones.

Verbal expressions are moderately compact; even at today's prices, the cost of letterpress composition is relatively readily borne. No editor forces the deletion of a sentence in a summary because it is a logical consequence of the sentences of the fuller exposition. Yet there is a standard that "the same information should not be given in both a graph and in a table". This is only in part because graphs and tables take much space and are expensive to set. It is also because the possibility of different purposes for two quantitative expressions of the same facts is neglected.

This appendix comments briefly on a few relevant examples. Its purpose is to stimulate the reader to think out some applications to his own work.

T1. COMPACT PRESENTATION

Tables must be relatively simple and full of white space if their messages are to be absorbed by the typical rapid reader. Graphs for similar purposes need to be simple and clearly labeled. These are precepts of broad application which we neglect at our peril.

Yet when quantitative information needs to be recorded, recorded only for those willing to dig, it can be compressed into text-like strings with great efficiency. John Hammersley was one of the pioneers of this (1954) when he recorded "full information" about a 4000-step self-avoiding random walk in 31 lines of Journal of the Royal Statistical Society text (pp. 31-32). There are readers who have read this paper through several times, and lectured on it to graduate students without learning how to decipher this compaction. The writer is one. But these readers know that they *can* recover the information if they need it. These lines are far more valuable as they now stand than they would be if the same space had been expended on a few additional sentences.

What are the prospects for such compactions elsewhere?

T2. COMPLETE PRESENTATION

Milton Friedman has reported (1957, p. 60, fn) his bitter experience in trying to make analytical use of government figures on consumer spending, where the policy of not giving figures based on few cases made otherwise valuable series wholly useless. To avoid danger to the naive or misguided, these tabulations were made useless to those who wished to give them specialized and serious study.

If this were an isolated instance, it would not deserve note. But it is not. And we dare not be surprised at its frequency. For it is a difficult typographical problem to combine either

(a) effective presentation of data to the rapidly scanning eye;

or

(b) protected presentation of data to the innocent;

with

(c) providing a record from which as much valuable information as possible can be recovered.

The task is not easy, yet an investigator who does not consider meeting both kinds of purposes is likely to be failing in his duty as a member of an on-going social institution.

T3. DWYER'S DEVICE

One of the simplest situations where it is desirable to provide for both the quick scan and the deep dig is in reporting frequency distributions. Unless the quantity distributed comes in neat little units (like number of children in a family), some grouping is inevitable, and heavy grouping is usually desired to save space.

A reasonable solution to this problem was suggested a number of years ago by Paul Dwyer (1942). Perhaps because of its location, this suggestion seems almost to have been lost. Yet it is simple and apparently effective.

Dwyer suggests that a grouped frequency distribution should wisely show

(a) the number of individuals in each cell;

(b) the sum of the values corresponding to these individuals; and

(c) the sum of squares of the values corresponding to these individuals.

He shows that this provides much more usable information when using heavily grouped frequency distributions than does (a) alone. (In a sense one becomes able to approximate with parabolic arcs rather than with horizontal steps.)

This proposal, which can be extended to more complex situations, seems to deserve much consideration.

T4. AN EXAMPLE

In seeking moderate persuasive examples of the use of nonclassical modes of expression we have turned to Volume 4 of ***Studies in Social Psychology in World War II*** ("Measurement and Prediction", Stouffer et al., 1950) in more than one instance. Our first simple

example (in Section E7) examined the 4-by-2 tables contrasting psychoneurotic patients with an army cross-section in a variety of questionnaire areas. And in the next appendix we shall examine one aspect of one further questionnaire area, area 16, psychosomatic complaints, for which no 4-by-2 table was given. Why?

Apparently because giving the "detailed data" in a graph (% frequency of occurrence for each number or score value) made it editorially impractical to provide the corresponding 4-by-2 table. Let us inquire into the sensibleness of this decision, even though inquiry may lead us into winding paths.

Why were the 4-by-2 tables given in detail in the first place? Presumably because they were relatively directly understandable by the target reader, and because they can be rather roughly compared with one another with modest ease. Such tables were given for 106 individual items and 16 summary scores, the 122 tables taking up 25 pages. Presumably both these tables and their comparisons were important to both author and editor. The omitted area involved the most different items and showed the greatest difference between neurotic patients and the cross-section of any of the areas. Yet comparisons involving it have to be made either by comparing a 4-by-2 table with a graph drawn for a different purpose or by reconstruction of a table from the graph.

If it were easy to read from the graph entries for a 4-by-2 table, the case would be less strong. But it is a tedious and delicate job to recover a possible set of entries. How can the saving of 1/123 of the space devoted to the 4-by-2 tables be justified, especially in a volume devoted to methodology?

As our examples demonstrate, it is now possible to handle the broader aspects of the information presented in each 4-by-2 table in terms of a few differences of logits. Thus these aspects could be easily condensed into a page or two of tables. In doing this, of course, there would be a loss of some of the numerical detail presented in the 4-by-2 tables. How could this be compactly recorded? How much space would it require?

One solution is easy to provide. If we use letters as break indicators and pass down columns in solid blocks, the % entries for the area 12 summary score, which is given in conventional form in Table 12 of Section E6, become "Scores a43b2c1d0e. Neurotic patients (%) a16b25c32d27e. Cross-section (%) a46b23c20d11." This would occupy about 1.3 lines of the large type used in the text of *Studies in Social Psychology in World War II*, but only about 1.0 line of the font used for table footnotes. This sort of condensation applied to area summary scores would lead to perhaps two lines of description, one of tabulation

of shifts in logits, and one of detailed numbers. For the individual items, the space required would be mainly that required to state the answers to which the %'s apply.

Consider next the graphs (Chart I on their page 501) which present distributions, both for psychoneurotics and for the cross-section, of each score according to three scoring systems. Their purpose was to show the substantial equivalence of simple and sophisticated scoring schemes. This purpose was accomplished to a limited extent, mainly by nongraphical means. The burden of the argument is carried by the similarity of certain %'s written prominently on the graphs. This similarity cannot be judged directly from the graphs, whose main virtue is to show that other critical score boundaries would not be obviously better. Something can be done to make the comparison directly appreciated graphically. The details are worked out in U3 below, and the results exhibited in Figures 14 and 15.

It is important to emphasize that examples were selected from "Measurement and Prediction," not because the source was technically poor, but because the source was technically good. It is only against a background of understanding of subject-matter, tender and loving care of data, and attention to exposition, that the detailed problems, difficulties, and solutions we have been discussing can be seen clearly and in silhouette. An example from a poor book would have failed to make its point.

U. MORE ABOUT MODES OF EXPRESSING COUNTED FRACTIONS

Chapter E gave considerable attention to three nonclassical modes of expressing fractions: anglits, normits, and logits. More attention there, though useful in itself, would have been too long a digression. In this appendix, then, we present three further examples (3A, 3B, and 3C), more details of tabulation (3D), a little about nature and behavior (3E), and some information on covariances between expressions of two fractions from a single table (3F).

U1. A SLIGHTLY MORE COMPLEX EXAMPLE

For a slightly more complex situation where a more reasonable mode of expression can be used to increase our understanding we turn again to Volume 4 of *Studies in Social Psychology in World War II*

(Stouffer et al. 1950), whose Table 2 on page 629 divides July 1945 and December 1945 separatees according to strength of plan to return to previous employer. Table 28 gives original %'s and the corresponding

Table 28

Plans to return to previous employer, by duration of previous employment in years. (From Volume 4 of *Studies in Social Psychology in World War II* (Stouffer et al. 1950, p. 629.)

	July Separatees				December Separatees			
	<1	1-2	2-5	>5	<1	1-2	2-5	>5
	(Individual % for each duration)							
Definite plans	12	22	31	49	18	35	47	64
Tentative plans	9	11	12	16	22	24	17	16
Considering returning	20	18	17	14	17	19	13	13
Not considering returning	59	49	40	21	43	22	23	17
	(Cumulative anglits for each duration)							
Definite plans	−.86	−.59	−.30	−.02	−.69	−.30	−.06	+.28
Tentative plans	−.62	−.35	−.14	+.30	−.20	+.18	+.28	+.64
Considering returning	−.18	+.02	+.20	+.62	+.14	+.59	+.57	+1.09
Not considering returning								

cumulative anglits, making use of the classification in terms of duration of previous employment. The two 3×4 tables which result may be dissected into (apparent) main effects and (apparent) interactions by the usual procedures of finding, and then subtracting, row means and column means. The results are shown in Tables 29 and 30.

The following conclusions appear to be supported by these last two tables:

(1) In terms of cumulative anglits, both tables show relatively small residuals; the approximate description in terms of main effects and grand means alone is quite effective.

(2) The large shift between July and December toward returning to the previous employer outweighs any other visible effects except the effect of duration of previous employment.

(3) The spread of opinion over the four-point scale is somewhat reduced in December as compared to July, the main effects of

Table 29

Results of dissecting the cumulative anglits of Table 28 for July separatees

Duration of previous employment

<1 year	1-2 years	2-5 years	>5 years
	(Undissected values)		
−.86	−.59	−.39	−.02
−.62	−.35	−.14	+.30
−.18	+.02	+.20	+.62

(residuals after dissection, bordered by main effects and grand mean)

−.01	+.02	+.02	−.02	−.30
−.03	+.00	+.01	+.04	−.04
+.04	+.00	−.02	−.01	+.33
−.37	−.14	+.06	+.47	−.17

Explanations:

(1) Any failure of rows or columns of residuals, or main effects to sum to zero is due to rounding of all answers to two decimal places.

(2) Each entry in undissected table is the sum of the corresponding 4 entries in the 4 dissected portions.

Example:

$$(-.86) = (-.01) + (-.30) + (-.37) + (-.17)$$

"break" being −.40 +.02, +.39 in place of −.30, +.04, +.33. (This may represent either a change in plans or a change in the way plans are described.)

(4) The described plans of those less than 1 year with previous employer are somewhat more different (show a somewhat greater shift against returning) in December than in July, the main effects being −.46 (vs. −.05, +.05, +.46) in place of −.37 (vs. −.14, +.06, +.47).

(5) The residuals for the middle break show slight trends for both groups, −.03, .00, .01, .04 in July and +.03, .00, .00, −.05 in December. The residuals for 1 to 2 and 2 to 5 years previous

Table 30

Results of dissecting the cumulative anglits of Table 28 for December separatees

Duration of previous employment

≤1 year	1-2 years	2-5 years	>5 years

(Undissected values)

−.69	−.30	−.06	+.28
−.20	+.18	+.28	+.64
+.14	+.59	+.57	1.09

(residuals after dissection, bordered by main effects and grand mean)

−.04	−.06	+.08	−.01	−.40
+.03	.00	.00	−.05	+.02
.00	+.04	−.08	+.03	+.39
−.46	−.05	+.05	+.46	+.21

Explanations:

(1) Any failure of rows or columns of residuals, or main effects to sum to zero is due to rounding of all answers to two decimal places.

(2) Each entry in undissected table is the sum of the corresponding 4 entries in the 4 dissected portions.

Example:

$$(-.86) = (-.01) + (-.30) + (-.37) + (-.17)$$

employment show trends in December, being −.06, .00, +.04 and +.08, .00, −.08, respectively. (Comparison for subgroups of separatees according to some other classification would be needed to indicate whether these effects deserve attention.)

The point at issue is again not the reality of such appearances, but whether or not they can be noticed and made the subject of reflection or study. It seems clear that the use of some mode of expression more compatible with the data than % was essential in bringing these appearances to the surface.

U2. UNORDERED FRACTIONS: ANOTHER EXAMPLE

Nonclassical modes of expression cannot only be useful in situations more complex than those of Sections E5 and E6, they can also be useful in still simpler situations. In the examples of Sections E6, E7 and U1, each group of units is divided into several fractions, fractions which are naturally arranged in an order. Logically simpler, though perhaps quantitatively harder to handle, is the situation where there are several fractions which do not appear to fall in any natural order.

An example of this revolves around data of Börje Hanssen (personal communication) on types of family names among heads of households in Strängnäs, Sweden. Table 31 shows the raw data, while

Table 31

Distribution of names of heads of households in Strängna, Sweden, according to type of name. (Data of Börje Hanssen)

Date	Total	Locality or physical characteristic	Occupational names	First name only	"—son" names	Bourgeois names
1652	191	16 (8.4%)	63 (33.0%)	7 (3.7%)	93 (48.7%)	12 (6.3%)
1689	236	7 (3.0%)	54 (22.9%)	28 (11.9%)	107 (45.4%)	40 (17.0%)
1727	194	1 (0.5%)	8 (4.1%)	12 (6.2%)	76 (39.2%)	97 (50.0%)
1740	224	—	—	4 (1.8%)	56 (25.0%)	164 (73.2%)
1813	378	—	—	—	22 (5.8%)	356 (94.2%)

Figures 11 to 13 show these data plotted against time according to various modes of expression, first as percentages, then as anglits, and lastly as logits.

To my eye, at least, the general "run" of the data improves steadily as we pass from the percentage mode, through the anglit mode, to the logit mode. In logits the curves run quite smoothly, suggesting reasonable extrapolation and interpolation, except for the unusually rapid conversion of "—son" names to bourgeois names between 1727 and 1740.

Whether or not such tail-stretching would be of real help in analyzing the changes of distribution of name types in Sweden cannot be settled by one trial. Only when data for a number of towns is available for comparison can we expect to find out.

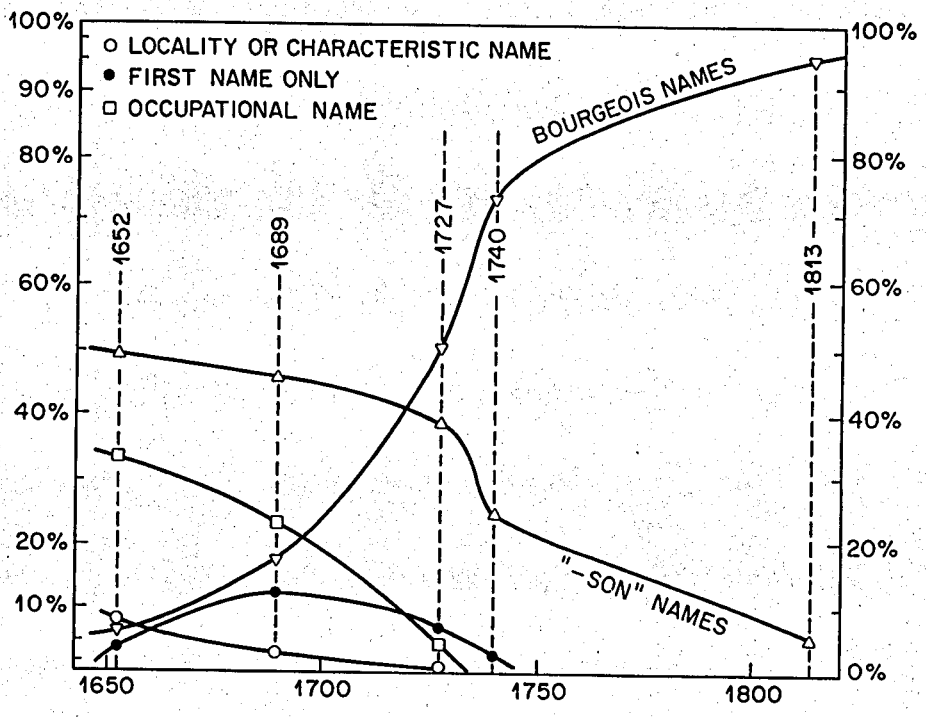


Figure 11. Kinds of name expressed in percentages.

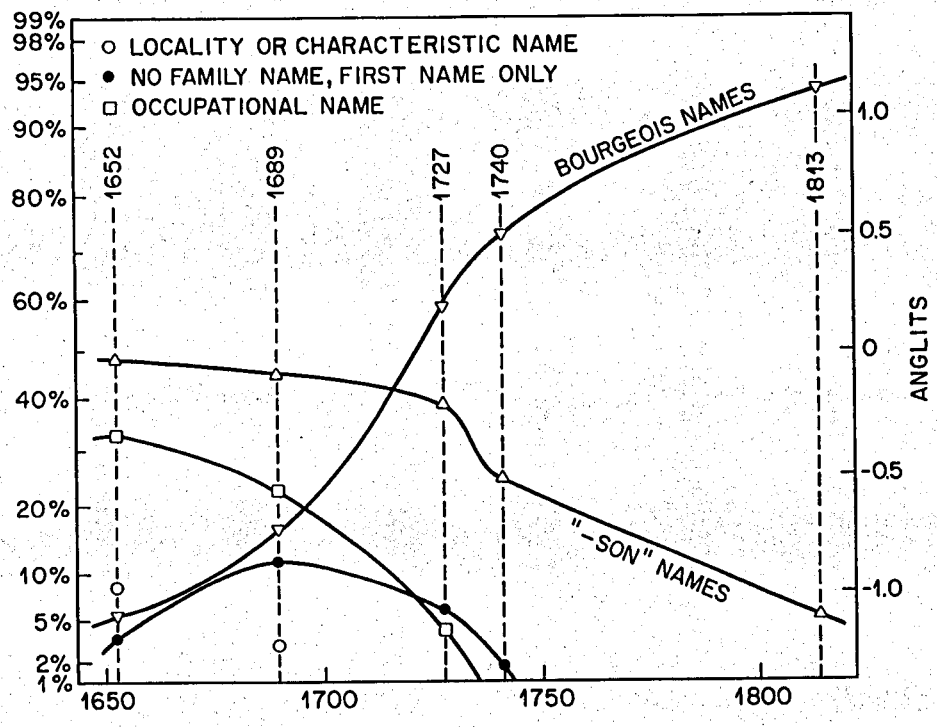


Figure 12. Kinds of name expressed in anglits

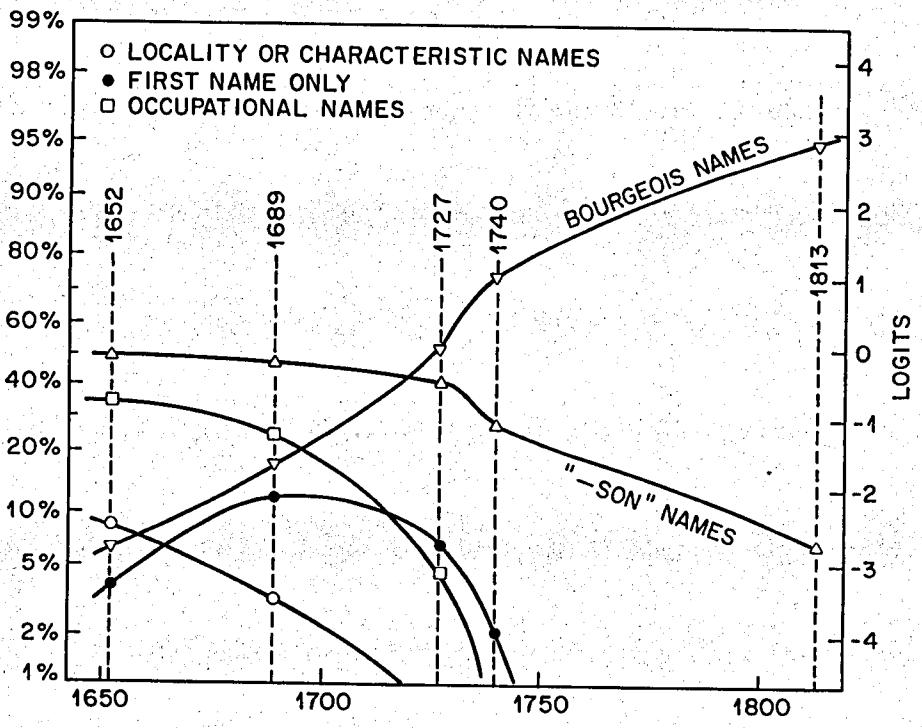


Figure 13. Kinds of name expressed in logits

U3. *AN EXAMPLE COMPARING DETAILED DISTRIBUTIONS*

A companion form of graph paper to those illustrated in Section E5 is Codex 41,453, 42,453, also designed by Berkson, which has normal cumulative probability scales both ways, thus allowing the plotting of one fraction against another with both expressed in nonclassical modes. In principle there could just as well be anglit-anglit paper and logit-logit paper as normit-normit paper, but these further kinds are not likely to become available until demand increases greatly. But with the tables of Sections E3 and D3, conversion of percentages into anglits or logits is easy, and the results can be treated by simple arithmetic, as well as being plotted against one another.

Chart 1 on page 501 of Volume 4 of *Studies in Social Psychology in World War II* (Stouffer et al. 1950) offers a good illustration of some of the possibilities. It presents distributions of summary scores on area 16, psychosomatic complaints, for the neurotic patients and Army cross-section samples which already appeared in Sections E6, T4 and U1. Three sorts of summary scores are considered: (i) simple dichotomous scores, where two-answer items are scored 0, 1, while three-answer items are scored 0, 0, 1 or 0, 1, 1; (ii) simple trichotomous scores where the three-answer items are scored 0, 1, 2, the two-answer items being treated as before; (iii) differential trichotomous scores where the weights are adjusted in terms of apparent predictive power.

Table 32 illustrates the numerical situation. (Note that the %'s given were read from a graph, and are undoubtedly full of small errors.) When the differences between the two distributions are examined for each of the modes of expression, it is clear that differences between anglits are badly lumped (low toward the tails) and are not likely to be as helpful or insight-generating as either of the other two.

The three weighting schemes are compared in terms of differences of logits in Figure 14, and in terms of differences of normits in Figure 15. The general conclusions to be drawn from either figure are the same, namely:

1) The differences between psychoneurotics and the cross-section were substantial but not strikingly large ($\approx 1.3\sigma$ for normits).

2) All three weighting schemes given generally quite similar results, as we should have expected.

3) Accordingly it is hard to identify any one weighting scheme as better than any other, although there may be a slight preference for simple trichotomous weights.

Table 32

Various presentations of the two frequency distributions for "simple dichotomous weights." (Based on top panel of Chart 1, page 501, of Stouffer et al., 1950.)

Score	Cumulative %		Half-logits		Differences Between Groups		
	Psych*	Cross*	Psych*	Cross*	Half-logits	Normits**	Anglits
0	5.6	0.0	−1.41	—			
1	14.0	0.6	−0.91	−2.6	1.7	1.14	.69
2	23.5	1.3	−.59	−2.15	1.55	1.21	.79
3	35.6	3.1	−.30	−1.72	1.42	1.19	.93
4	48.9	5.3	−.02	−1.44	1.42	1.28	1.20
5	60.4	8.3	.21	−1.20	1.41	1.32	1.32
6	68.6	11.9	.39	−1.00	1.39	1.33	1.24
7	77.0	16.3	.60	−.82	1.42	1.36	1.31
8	83.9	21.4	.82	−.65	1.47	1.41	1.35
9	89.4	28.6	1.07	−.46	1.53	1.45	1.35
10	93.3	36.2	1.31	−.29	1.60	1.49	1.34
11	95.5	45.7	1.52	−.09	1.61	1.44	1.23
12	97.2	57.3	1.78	+.15	1.63	1.38	1.09
13	99.2	71.6	2.41	.47	1.94	1.47	.94
14	100.0	87.7	∞	.47			
15	100.0	100.	∞	∞			

* Psych = 563 psychoneurotic patients in Army hospitals
Cross = 3,501 white enlisted men without overseas service

** Special normits with 0.798 · · · multiplier

4) In either normit or logit terms, the differences were greater for the end of the distribution (to the right in the plots) corresponding to a high psychoneurosis score — one extreme end showing a ratio of perhaps 3 to 2 compared to the other.

Contrasting the figures, we also see that:

5) As was inevitable, differences in logits near either end are enhanced by comparison with differences of normits and of near-center behavior. The overall impression is concave upward rather than nearly straight.

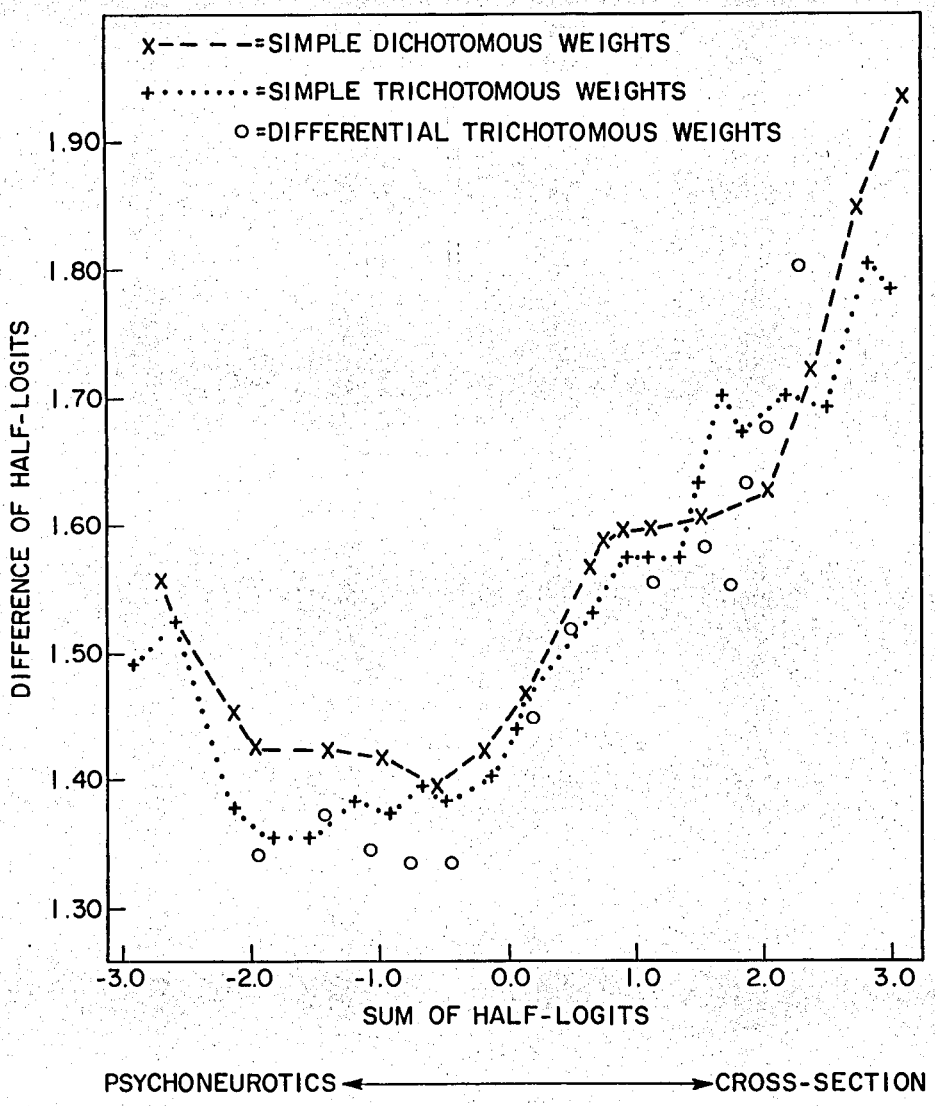


Figure 14. Comparison in terms of logits

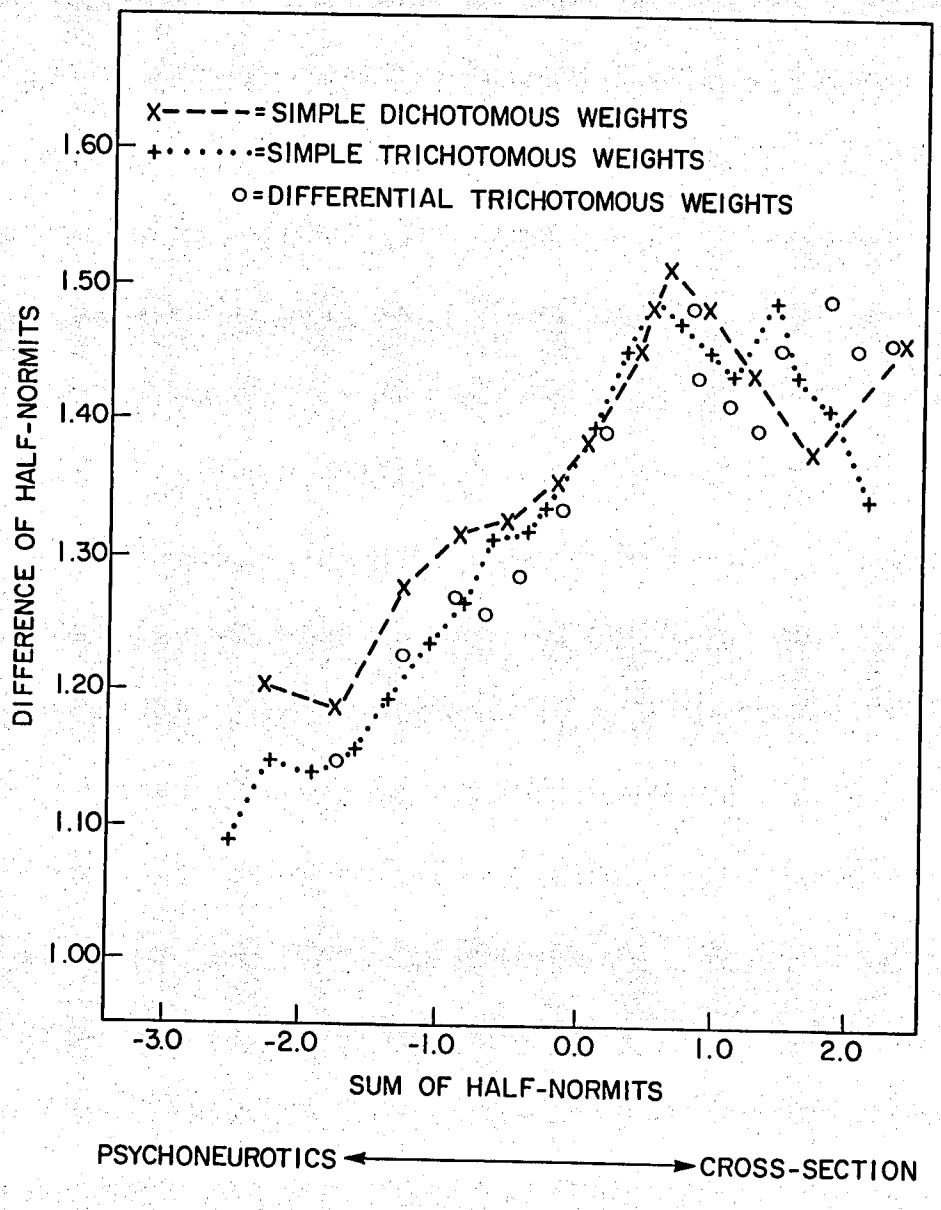


Figure 15. Comparison in terms of normits

U4. FURTHER TABLES FOR NONCLASSICAL MODES

Table 10 in Section E6 is a simple table by conventional standards. It is entered with a fraction expressed as a percentage, and, frequently after some simple interpolation, it provides two-decimal-place values of the mode of expression chosen. It may not be easy to see how the process can be made simpler.

Tables of logarithms are usually given to 5 or more decimal places, sometimes to only 4. If they are to be used for the classical purposes of logarithm tables (conversion of multiplication into addition and division

Table 33

Critical table of one-decimal logarithms

Leading nonzero digits of argument	First decimal of logarithm
(890 · · ·)	
	0
112- · · ·	
	1
141- · · ·	
	2
177- · · ·	
	3
2239 · · ·	
	4
2819 · · ·	
	5
354- · · ·	
	6
446- · · ·	
	7
562- · · ·	
	8
708- · · ·	
	9
890- · · ·	

into subtraction so as to ease the pain of arithmetic without machine aid), these precisions are quite natural, and even necessary. But when a logarithm table is to be used to change the mode of expression of a rather crudely measured quantity, there is need for far less precision. And using fewer decimals will ease the arithmetic.

Table 33 contains ten numerical values, one repeated. It is a *critical table* of one-decimal logarithms (to the base 10). Its use may be illustrated as follows: Given 33.725; to find its one-decimal logarithm: Note first that 33.725 is at least 10, and less than 100, so the integer part of the logarithm is 1; referring to Table 33, 33725 falls between 2819. and 354-. hence the first decimal is .5 and the whole (one decimal) logarithm is 1.5. Similarly, 0.0739, is at least 0.01 and less than 0.1, hence the integer part of its logarithm is −2, while, in Table 33, 739 lies between 708 and 890, so that the decimal part is .9 and the whole logarithm is $-2 + 0.9 = 1.1$. Note that no interpolation is ever needed.

Critical tables can always be easily constructed if the results are not required to too high precision. A critical table for two-decimal logarithms is very useful, but one for three or more decimals would be far less convenient. Once we come to a critical table for two-decimal logarithms, with its 100 entries, it is no longer desirable to have the entries in a single column. Table 34 presents a critical table for two-decimal logarithms in a square array. It is to be used by reading down

Table 34

Two-decimal critical table of common logarithms

	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09	
.0	9886	1012	1035	1059	1084	1109	1135	1161	1189	1216	1245 .0
.1	1245	1274	1303	1334	1365	1396	1429	1462	1496	1531	1567 .1
.2	1567	1603	1641	1679	1718	1758	1799	1844	1884	1928	1972 .2
.3	1972	2018	2065	2113	2163	2213	2265	2317	2371	2427	2483 .3
.4	2483	2541	2600	2661	2723	2786	2851	2917	2985	3055	3126 .4
.5	3126	3199	3273	3350	3428	3508	3589	3673	3758	3846	3936 .5
.6	3936	4027	4121	4217	4315	4416	4519	4624	4732	4842	4955 .6
.7	4955	5070	5188	5309	5433	5559	5689	5821	5957	6095	6237 .7
.8	6237	6383	6531	6683	6839	6998	7161	7328	7499	7674	7852 .8
.9	7852	8035	8222	8414	8610	8810	9016	9226	9441	9661	9886 .9
	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09	

the first column to locate the broad gap in which the given value falls, which fixes the first digit, and then reading horizontally (in the line above this gap) to locate the narrow gap in which the given value falls, thus locating the second digit. If, for example, the number is 3873, we find 3126 above 3936 in the left-hand column of the body of the table, and then scan the row beginning with 3126 (the row associated with .5) to find 3846 followed by 3936. This latter gap is associated with .5 (in the line) and .09 (gap between columns), so that the answer is 0.59 plus the characteristic, which yields 3.59.

Tables 35, 36 and 37 provide two-decimal critical tables for the modes of Section E4; anglits, matched normits and matched logits. As is pointed out in Section S3, this precision is adequate for simple random samples of several thousand, and, indeed, probably adequate for almost all samples actually available.

Theoretical work is sometimes facilitated by additional precision. Accordingly, Table 38 gives values of anglits, (unmodified) normits, and (unmodified) logits to four decimal places and the round %'s used in Table 10.

U5. SOME PROPERTIES OF THE THREE MODES

Not because they are important to our present concerns, but only because they may be curiosity-allaying or intuition-increasing, we present here a small amount of information about the mathematical definitions and statistical properties of anglits, normits and logits. (This material is in an appendix in the hope that the less mathematically-minded reader will skip it.)

The gentlest in tail-stretching of these three modes is represented by the use of anglits, of angles θ satisfying

$$\sin^2 \theta = \text{(fraction observed)}.$$

(Varied choices of unit for θ are used; degrees or radians are used, the values of θ are sometimes doubled and sometimes not, and a constant may or may not be subtracted to make zero the anglit corresponding to 50%. These choices are not essentially different; almost all further analyses will lead to the same results whichever *one* be used. Danger and confusion is only possible when two or more of these choices are confused and combined. We have used here the choice corresponding to the graph paper which was illustrated (32,452) in which the choices are (i) to use radians, (ii) to double, (iii) to subtract the constant.)

This mode (often referred to as the angular transformation or the arc-sine transformation) has the following interesting properties:

(1) There are "ends" to the scale: for our choices all anglits will lie between -1.571 and $+1.571$. (The exact ends are at $\pm\pi/2$.)

(2) Near these ends, the deviation of the anglit from the end value is approximately twice the square root of the smaller observed fraction.

(3) In simple random sampling, the variance of the anglit is quite closely 1/(total size of sample).

The middle one of the three modes, so far as tail-stretching is concerned, is represented by the use of normits or probits, that is, by the use of expressions connected with the cumulative normal distribution. The use of this mode is sometimes justified by hypotheses involving an underlying continuous scale, a threshold point on the underlying scale, and normally distributed perturbations which give particular situations probability, rather than certainty, of appearing to fall on one side of the threshold rather than the other. Such justifications can be quite frequently helpful, and are even sometimes close to being correct, although it is often *very important* that other plausible structures also lead to the use of normits or probits. The best single justification for this mode, as for all others, is empirical. If it demonstrably works, fine. If it is demonstrably better than competing modes, finer still. The data is the final test.

Gory details of definition need not detain us, but for the record, and because some may care, we note that the more usual representatives of this mode are normits and probits, where

$$x = \text{normit of } p$$

means

$$p = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2}\, du$$

and

$$\text{probit of } p = 5 + (\text{normit of } p).$$

Table 35

Critical table from fractions expressed as % to anglits. (For values ⩽ 50%.)

	.00	−.01	−.02	−.03	−.04	−.05	−.06	−.07	−.08	−.09	
−.0	50.25	49.75	49.25	48.75	48.25	47.75	47.25	46.75	46.25	45.76	45.26	−.0
−.1	45.26	44.76	44.26	43.77	43.27	42.78	42.28	41.79	41.29	40.80	40.31	−.1
−.2	40.31	39.82	39.33	38.84	38.36	37.87	37.39	36.90	36.42	35.94	35.46	−.2
−.3	35.46	34.99	34.51	34.03	33.56	33.09	32.62	32.15	31.69	31.22	30.76	−.3
−.4	30.76	30.30	29.84	29.38	28.93	28.48	28.03	27.58	27.13	26.69	26.25	−.4
−.5	26.25	25.81	25.37	24.94	24.51	24.08	23.65	23.23	22.81	22.39	21.97	−.5
−.6	21.97	21.56	21.15	20.75	20.34	19.94	19.54	19.15	18.76	18.37	17.98	−.6
−.7	17.98	17.60	17.22	16.84	16.47	16.10	15.74	15.37	15.01	14.66	14.31	−.7
−.8	14.31	13.96	13.61	13.27	12.94	12.60	12.27	11.95	11.62	11.30	10.99	−.8
−.9	10.99	10.68	10.37	10.07	9.77	9.48	9.18	8.90	8.61	8.34	8.06	−.9
−1.0	8.06	7.79	7.53	7.26	7.01	6.75	6.50	6.26	6.02	5.78	5.55	−1.0
−1.1	5.55	5.33	5.10	4.89	4.67	4.46	4.26	4.06	3.87	3.68	3.49	−1.1
−1.2	3.49	3.31	3.13	2.96	2.79	2.63	2.47	2.32	2.17	2.03	1.89	−1.2
−1.3	1.89	1.76	1.63	1.50	1.38	1.27	1.16	1.06	0.96	0.86	0.77	−1.3
−1.4	0.77	0.69	0.61	0.53	0.46	0.40	0.33	0.28	0.23	0.18	0.144	−1.4
−1.5	0.144	0.108	0.078	0.052	0.032	0.017	0.006	0.001	0.000			−1.5
	.00	−.01	−.02	−.03	−.04	−.05	−.06	−.07	−.08	−.09		

Table 35 (Cont'd)

Critical table from fractions expressed as % to anglits. (For values ⩾ 50%.)

	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09	
.0	49.75	50.25	50.75	51.25	51.75	52.25	52.75	53.25	53.75	54.24	54.74	.0
.1	54.74	55.24	55.74	56.23	56.73	57.22	57.72	58.21	58.71	59.20	59.69	.1
.2	59.69	60.18	60.67	61.16	61.64	62.13	62.61	63.10	63.58	64.06	64.54	.3
.3	64.54	65.01	65.49	65.97	66.44	66.91	67.38	67.85	68.31	68.78	69.24	.3
.4	69.24	69.70	70.16	70.62	71.07	71.52	71.97	72.42	72.87	73.31	73.75	.4
.5	73.75	74.19	74.63	75.06	75.49	75.92	76.35	76.77	77.19	77.61	78.03	.5
.6	78.03	78.44	78.85	79.25	79.66	80.06	80.46	80.85	81.24	81.63	82.02	.6
.7	82.02	82.40	82.78	83.16	83.53	83.90	84.26	84.63	84.99	85.34	85.69	.7
.8	85.69	86.04	86.39	86.73	87.06	87.40	87.73	88.05	88.38	88.70	89.01	.8
.9	89.01	89.32	89.63	89.93	90.23	90.52	90.82	91.10	91.39	91.66	91.94	.9
1.0	91.94	92.21	92.47	92.74	92.99	93.25	93.50	93.74	93.98	94.22	94.45	1.0
1.1	94.45	94.67	94.90	95.11	95.33	95.54	95.74	95.94	96.13	96.32	96.51	1.1
1.2	96.51	96.69	96.87	97.04	97.21	97.37	97.53	97.68	97.83	97.97	98.11	1.2
1.3	98.11	98.24	98.37	98.50	98.62	98.73	98.84	98.94	99.04	99.14	99.23	1.3
1.4	99.23	99.31	99.39	99.47	99.54	99.60	99.67	99.72	99.77	99.82	99.856	1.4
1.5	99.856	99.892	99.922	99.948	99.968	99.983	99.994	99.999	100.000			1.5
	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09		

Table 36

Critical table from fractions expressed as % to fractions expressed as (matched, modified) normits. (For values ⩽ 50%.)

	.00	−.01	−.02	−.03	−.04	−.05	−.06	−.07	−.08	−.09	
−.0	50.25	49.75	49.25	48.75	48.25	47.75	47.25	46.75	46.26	45.76	45.26	−.0
−.1	45.26	44.77	44.27	43.78	43.28	42.79	42.30	41.81	41.32	40.83	40.35	−.1
−.2	40.35	39.86	39.38	38.90	38.42	37.94	37.46	36.99	36.52	36.05	35.58	−.2
−.3	35.58	35.11	34.65	34.19	33.73	33.27	32.82	32.37	31.92	31.47	31.03	−.3
−.4	31.03	30.59	30.15	29.71	29.28	28.85	28.43	28.00	27.58	27.16	26.75	−.4
−.5	26.75	26.34	25.93	25.53	25.13	24.73	24.33	23.94	23.56	23.17	22.79	−.5
−.6	22.79	22.41	22.04	21.67	21.31	20.94	20.58	20.23	19.88	19.53	19.19	−.6
−.7	19.19	18.85	18.51	18.18	17.85	17.52	17.20	16.88	16.57	16.26	15.95	−.7
−.8	15.95	15.65	15.35	15.06	14.77	14.48	14.20	13.92	13.64	13.37	13.10	−.8
−.9	13.10	12.83	12.57	12.32	12.06	11.81	11.57	11.32	11.09	10.85	10.62	−.9
−1.0	10.62	10.39	10.17	9.95	9.73	9.51	9.30	9.10	8.89	8.69	8.50	−1.0
−1.1	8.50	8.30	8.11	7.93	7.74	7.56	7.39	7.21	7.04	6.87	6.71	−1.1
−1.2	6.71	6.55	6.39	6.24	6.08	5.93	5.79	5.64	5.50	5.36	5.23	−1.2
−1.3	5.23	5.10	4.97	4.84	4.71	4.59	4.47	4.36	4.24	4.13	4.02	−1.3
−1.4	4.02	3.91	3.81	3.71	3.60	3.51	3.41	3.32	3.23	3.14	3.05	−1.4
−1.5	3.05	2.96	2.88	2.80	2.72	2.64	2.57	2.49	2.42	2.35	2.28	−1.5
	.00	−.01	−.02	−.03	−.04	−.05	−.06	−.07	−.08	−.09	

Table 36 (Cont'd)

Critical table from fractions expressed as % to fractions expressed as (matched, modified) normits. (For values $\leqslant 50\% -$ cont'd.)

	.00	−.01	−.02	−.03	−.04	−.05	−.06	−.07	−.08	−.09		
−1.6	2.28	2.21	2.15	2.08	2.02	1.96	1.90	1.85	1.79	1.74	1.68	−1.6
−1.7	1.68	1.63	1.58	1.53	1.48	1.44	1.39	1.35	1.31	1.26	1.22	−1.7
−1.8	1.22	1.18	1.15	1.11	1.07	1.04	1.00	0.97	0.94	0.91	0.877	−1.8
−1.9	0.877	0.848	0.820	0.792	0.765	0.739	0.714	0.689	0.666	0.643	0.620	−1.9
−2.0	0.620	0.599	0.578	0.558	0.538	0.519	0.500	0.483	0.465	0.449	0.432	−2.0
−2.1	0.432	0.417	0.402	0.387	0.373	0.359	0.346	0.333	0.321	0.309	0.297	−2.1
−2.2	0.297	0.286	0.275	0.265	0.255	0.245	0.236	0.226	0.218	0.209	0.201	−2.2
−2.3	0.201	0.193	0.186	0.178	0.171	0.165	0.158	0.152	0.146	0.140	0.134	−2.3
−2.4	0.134	0.129	0.124	0.119	0.114	0.109	0.105	0.100	0.096	0.092	0.088	−2.4
−2.5	0.088	0.085	0.081	0.078	0.074	0.071	0.068	0.065	0.062	0.060	0.057	−2.5
−2.6	0.057	0.055	0.052	0.050	0.048	0.046	0.044	0.042	0.040	0.038	0.037	−2.6
−2.7	0.037	0.035	0.033	0.032	0.030	0.029	0.028	0.026	0.025	0.024	0.023	−2.7
−2.8	0.023	0.022	0.021	0.020	0.019	0.018	0.018	0.016	0.016	0.015	0.014	−2.8
−2.9	0.014	0.014	0.013	0.012	0.012	0.011	0.011	0.010	0.010	0.009	0.009	−2.9
−3.0	0.009	0.008										−3.0

.00 −.01 −.02 −.03 −.04 −.05 −.06 −.07 −.08 −.09

Table 36 (Cont'd)

Critical table from fractions expressed as % to fractions expressed as (matched, modified) normits. (For values ⩾ 50%.)

	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09	
.0	49.75	50.25	50.75	51.25	51.75	52.25	52.75	53.25	53.74	54.24	54.74	.0
.1	54.74	55.23	55.73	56.22	56.72	57.21	57.70	58.19	58.68	59.17	59.65	.1
.2	59.65	60.14	60.62	61.10	61.58	62.06	62.54	63.01	63.48	63.95	64.42	.2
.3	64.42	64.89	65.35	65.81	66.27	66.73	67.18	67.63	68.08	68.53	68.97	.3
.4	68.97	69.41	69.85	70.29	70.72	71.15	71.57	72.00	72.42	72.84	73.25	.4
.5	73.25	73.66	74.07	74.47	74.87	75.27	75.67	76.06	76.44	76.83	77.21	.5
.6	77.21	77.59	77.96	78.33	78.69	79.06	79.42	79.77	80.12	80.47	80.81	.6
.7	80.81	81.15	81.49	81.82	82.15	82.48	82.80	83.12	83.43	83.74	84.05	.7
.8	84.05	84.35	84.65	84.94	85.23	85.52	85.80	86.08	86.36	86.63	86.90	.8
.9	86.90	87.17	87.43	87.68	87.94	88.19	88.43	88.68	88.91	89.15	89.38	.9
1.0	89.38	89.61	89.83	90.05	90.27	90.49	90.70	90.90	91.11	91.31	91.50	1.0
1.1	91.50	91.70	91.89	92.07	92.26	92.44	92.61	92.79	92.96	93.13	93.29	1.1
1.2	93.29	93.45	93.61	93.76	93.92	94.07	94.21	94.36	94.50	94.64	94.77	1.2
1.3	94.77	94.90	95.03	95.16	95.29	95.41	95.53	95.64	95.76	95.87	95.98	1.3
1.4	95.98	96.09	96.19	96.29	96.40	96.49	96.59	96.68	96.77	96.86	96.95	1.4
1.5	96.95	97.04	97.12	97.20	97.28	97.36	97.43	97.51	97.58	97.65	97.72	1.5

	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09

Table 36 (Cont'd)

Critical table from fractions expressed as % to fractions expressed as (matched, modified) normits. (For values $\geqslant 50\% -$ cont'd.)

	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09	
1.6	97.72	97.79	97.85	97.92	97.98	98.04	98.10	98.15	98.21	98.26	98.32	1.6
1.7	98.32	98.37	98.42	98.47	98.52	98.56	98.61	98.65	98.69	98.74	98.78	1.7
1.8	98.78	98.82	98.85	98.89	98.93	98.96	99.00	99.03	99.06	99.09	99.123	1.8
1.9	99.123	99.152	99.180	99.208	99.235	99.261	99.286	99.311	99.334	99.357	99.380	1.9
2.0	99.380	99.401	99.422	99.442	99.462	99.481	99.500	99.517	99.535	99.551	99.568	2.0
2.1	99.568	99.583	99.598	99.613	99.627	99.641	99.654	99.667	99.679	99.691	99.703	2.1
2.2	99.703	99.714	99.725	99.735	99.745	99.755	99.764	99.774	99.782	99.791	99.799	2.2
2.3	99.799	99.807	99.814	99.822	99.829	99.835	99.842	99.848	99.854	99.860	99.866	2.3
2.4	99.866	99.871	99.876	99.881	99.886	99.891	99.895	99.900	99.904	99.908	99.912	2.4
2.5	99.912	99.915	99.919	99.922	99.926	99.929	99.932	99.935	99.938	99.940	99.943	2.5
2.6	99.943	99.945	99.948	99.950	99.952	99.954	99.956	99.958	99.960	99.962	99.963	2.6
2.7	99.963	99.965	99.967	99.968	99.970	99.971	99.972	99.974	99.975	99.976	99.977	2.7
2.8	99.977	99.978	99.979	99.980	99.981	99.982	99.982	99.984	99.984	99.985	99.986	2.8
2.9	99.986	99.986	99.987	99.988	99.988	99.989	99.989	99.990	99.990	99.991	99.991	2.9
3.0	99.991	99.992										3.0
	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09		

Table 37

Critical table from fractions expressed as % to fractions expressed as (matched, half-) logits. (For values ⩽ 50%.)

	.00	−.01	−.02	−.03	−.04	−.05	−.06	−.07	−.08	−.09	
−.0	50.25	49.75	49.25	48.75	48.25	47.75	47.25	46.75	46.26	45.76	45.26	−.0
−.1	45.26	44.77	44.28	43.78	43.29	42.80	42.31	41.82	41.34	40.85	40.37	−.1
−.2	40.37	39.89	39.41	38.94	38.46	37.99	37.52	37.05	36.59	36.12	35.66	−.2
−.3	35.66	35.21	34.75	34.30	33.85	33.40	32.96	32.52	32.08	31.65	31.22	−.3
−.4	31.22	30.79	30.36	29.94	29.53	29.11	28.70	28.29	27.89	27.49	27.09	−.4
−.5	27.09	26.70	26.31	25.92	25.54	25.16	24.79	24.42	24.05	23.69	23.33	−.5
−.6	23.33	22.97	22.62	22.27	21.93	21.59	21.25	20.92	20.59	20.26	19.94	−.6
−.7	19.94	19.62	19.31	19.00	18.69	18.39	18.09	17.80	17.51	17.22	16.94	−.7
−.8	16.94	16.66	16.38	16.11	15.84	15.58	15.32	15.06	14.80	14.55	14.31	−.8
−.9	14.31	14.06	13.82	13.59	13.35	13.12	12.90	12.68	12.46	12.24	12.03	−.9
−1.0	12.03	11.82	11.61	11.41	11.20	11.01	10.81	10.62	10.43	10.25	10.07	−1.0
−1.1	10.07	9.89	9.71	9.53	9.36	9.20	9.03	8.87	8.71	8.55	8.39	−1.1
−1.2	8.39	8.24	8.09	7.94	7.80	7.66	7.52	7.38	7.24	7.11	6.98	−1.2
−1.3	6.98	6.85	6.72	6.60	6.48	6.36	6.24	6.12	6.01	5.90	5.79	−1.3
−1.4	5.79	5.68	5.57	5.47	5.37	5.27	5.17	5.07	4.97	4.88	4.79	−1.4
−1.5	4.79	4.70	4.61	4.52	4.44	4.35	4.27	4.19	4.11	4.03	3.95	−1.5
−1.6	3.95	3.88	3.81	3.73	3.66	3.59	3.52	3.46	3.39	3.32	3.26	−1.6
−1.7	3.26	3.20	3.14	3.08	3.02	2.96	2.90	2.85	2.79	2.74	2.69	−1.7
−1.8	2.69	2.63	2.58	2.53	2.48	2.44	2.39	2.34	2.30	2.25	2.21	−1.8
−1.9	2.21	2.17	2.12	2.08	2.04	2.00	1.96	1.93	1.89	1.85	1.82	−1.9
−2.0	1.82	1.78	1.75	1.71	1.68	1.65	1.61	1.58	1.51	1.52	1.49	−2.0
	.00	−.01	−.02	−.03	−.04	−.05	−.06	−.07	−.08	−.09	

Table 37 (Cont'd)

Critical table from fractions expressed as % to fractions expressed as (matched, half-) logits. (For values ⩽ 50% − cont'd.)

	.00	−.01	−.02	−.03	−.04	−.05	−.06	−.07	−.08	−.09	
−2.1	1.49	1.46	1.43	1.41	1.38	1.35	1.33	1.30	1.27	1.25	1.22	−2.1
−2.2	1.22	1.20	1.18	1.15	1.13	1.11	1.09	1.07	1.05	1.03	1.01	−2.2
−2.3	1.01	0.99	0.97	0.95	0.93	0.91	0.89	0.87	0.86	0.84	0.824	−2.3
−2.4	0.824	0.808	0.792	0.777	0.761	0.747	0.732	0.717	0.703	0.690	0.676	−2.4
−2.5	0.676	0.663	0.650	0.637	0.624	0.612	0.600	0.588	0.577	0.565	0.554	−2.5
−2.6	0.554	0.543	0.533	0.522	0.512	0.502	0.492	0.482	0.473	0.463	0.454	−2.6
−2.7	0.454	0.445	0.436	0.428	0.419	0.411	0.403	0.395	0.387	0.380	0.372	−2.7
−2.8	0.372	0.365	0.358	0.351	0.344	0.337	0.330	0.324	0.317	0.311	0.305	−2.8
−2.9	0.305	0.299	0.293	0.287	0.281	0.276	0.270	0.265	0.260	0.255	0.250	−2.9
−3.0	0.250	0.245	0.240	0.235	0.231	0.226	0.222	0.217	0.213	0.209	0.205	−3.0
−3.1	0.205	0.201	0.197	0.193	0.189	0.185	0.181	0.178	0.174	0.171	0.168	−3.1
−3.2	0.168	0.164	0.161	0.158	0.155	0.152	0.149	0.146	0.143	0.140	0.137	−3.2
−3.3	0.137	0.135	0.132	0.129	0.127	0.124	0.122	0.119	0.117	0.115	0.112	−3.3
−3.4	0.112	0.110	0.108	0.106	0.104	0.102	0.100	0.098	0.096	0.094	0.092	−3.4
−3.5	0.092	0.090	0.088	0.087	0.085	0.083	0.082	0.080	0.078	0.077	0.075	−3.5
−3.6	0.075	0.074	0.072	0.071	0.070	0.068	0.067	0.066	0.064	0.063	0.062	−3.6
−3.7	0.062	0.060	0.059	0.058	0.057	0.056	0.055	0.054	0.053	0.052	0.051	−3.7
−3.8	0.051	0.050	0.049	0.048	0.047	0.046	0.045	0.044	0.043	0.042	0.041	−3.8
−3.9	0.041	0.041	0.040	0.039	0.038	0.037	0.037	0.036	0.035	0.035	0.034	−3.9
−4.0	0.034	0.033										−4.0
	.00	−.01	−.02	−.03	−.04	−.05	−.06	−.07	−.08	−.09		

Table 37 (Cont'd)

Critical table from fractions expressed as % to fractions expressed as (matched, half-) logits. (For values ⩾ 50%.)

	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09	
.0	49.75	50.25	50.75	51.25	51.75	52.25	52.75	53.25	53.74	54.24	54.74	.0
.1	54.74	55.23	55.72	56.22	56.71	57.20	57.69	58.18	58.66	59.15	59.63	.1
.2	59.63	60.11	60.59	61.06	61.54	62.01	62.48	62.95	63.41	63.88	64.34	.2
.3	64.34	64.79	65.25	65.70	66.15	66.60	67.04	67.48	67.92	68.35	68.78	.3
.4	68.78	69.21	69.64	70.06	70.47	70.89	71.30	71.71	72.11	72.51	72.91	.4
.5	72.91	73.30	73.69	74.08	74.46	74.84	75.21	75.58	75.95	76.31	76.67	.5
.6	76.67	77.03	77.38	77.73	78.07	78.41	78.75	79.08	79.41	79.74	80.06	.6
.7	80.06	80.38	80.69	81.00	81.31	81.61	81.91	82.20	82.49	82.78	83.06	.7
.8	83.06	83.34	83.62	83.89	84.16	84.42	84.68	84.94	85.20	85.45	85.69	.8
.9	85.69	85.94	86.18	86.41	86.65	86.88	87.10	87.32	87.54	87.76	87.97	.9
1.0	87.97	88.18	88.39	88.59	88.80	88.99	89.19	89.38	89.57	89.75	89.93	1.0
1.1	89.93	90.11	90.29	90.47	90.64	90.80	90.97	91.13	91.29	91.45	91.61	1.1
1.2	91.61	91.76	91.91	92.06	92.20	92.34	92.48	92.62	92.76	92.89	93.02	1.2
1.3	93.02	93.15	93.28	93.40	93.52	93.64	93.76	93.88	93.99	94.10	94.21	1.3
1.4	94.21	94.32	94.43	94.53	94.63	94.73	94.83	94.93	95.03	95.12	95.21	1.4
1.5	95.21	95.30	95.39	95.48	95.56	95.65	95.73	95.81	95.89	95.97	96.05	1.5
1.6	96.05	96.12	96.19	96.27	96.34	96.41	96.48	96.54	96.61	96.68	96.74	1.6
1.7	96.74	96.80	96.86	96.92	96.98	97.04	97.10	97.15	97.21	97.26	97.31	1.7
1.8	97.31	97.37	97.42	97.47	97.52	97.56	97.61	97.66	97.70	97.75	97.79	1.8
1.9	97.79	97.83	97.88	97.92	97.96	98.00	98.04	98.07	98.11	98.15	98.18	1.9
2.0	98.18	98.22	98.25	98.29	98.32	98.35	98.39	98.42	98.49	98.48	98.51	2.0

.00 .01 .02 .03 .04 .05 .06 .07 .08 .09

Table 37 (Cont'd)

Critical table from fractions expressed as % to fractions expressed as (matched, half-) logits. (For values ⩾ 50% – cont'd.)

	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09	
2.1	98.51	98.54	98.57	98.59	98.62	98.65	98.67	98.70	98.73	98.75	98.78	2.1
2.2	98.78	98.80	98.82	98.85	98.87	98.89	98.91	98.93	98.95	98.97	98.99	2.2
2.3	98.99	99.01	99.03	99.05	99.07	99.09	99.11	99.13	99.14	99.16	99.176	2.3
2.4	99.176	99.192	99.208	99.223	99.239	99.253	99.268	99.283	99.297	99.310	99.324	2.4
2.5	99.324	99.337	99.350	99.363	99.376	99.388	99.400	99.412	99.423	99.435	99.446	2.5
2.6	99.446	99.457	99.467	99.478	99.488	99.498	99.508	99.518	99.527	99.537	99.546	2.6
2.7	99.546	99.555	99.564	99.572	99.581	99.589	99.597	99.605	99.613	99.620	99.628	2.7
2.8	99.628	99.635	99.642	99.649	99.656	99.663	99.670	99.676	99.683	99.689	99.695	2.8
2.9	99.695	99.701	99.707	99.713	99.719	99.724	99.730	99.735	99.740	99.745	99.750	2.9
3.0	99.750	99.755	99.760	99.765	99.769	99.774	99.778	99.783	99.787	99.791	99.795	3.0
3.1	99.795	99.799	99.803	99.807	99.811	99.815	99.819	99.822	99.826	99.829	99.832	3.1
3.2	99.832	99.836	99.839	99.842	99.845	99.848	99.851	99.854	99.857	99.860	99.863	3.2
3.3	99.863	99.865	99.868	99.871	99.873	99.876	99.878	99.881	99.883	99.885	99.888	3.3
3.4	99.888	99.890	99.892	99.894	99.896	99.898	99.900	99.902	99.904	99.906	99.908	3.4
3.5	99.908	99.910	99.912	99.913	99.915	99.917	99.918	99.920	99.922	99.923	99.925	3.5
3.6	99.925	99.926	99.928	99.929	99.930	99.932	99.933	99.934	99.936	99.937	99.938	3.6
3.7	99.938	99.940	99.941	99.942	99.943	99.944	99.945	99.946	99.947	99.948	99.949	3.7
3.8	99.949	99.950	99.951	99.952	99.953	99.954	99.955	99.956	99.957	99.958	99.959	3.8
3.9	99.959	99.959	99.960	99.961	99.962	99.963	99.963	99.964	99.965	99.965	99.966	3.9
4.0	99.966	99.967										4.0
	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09		

Table 38

Values of anglits, normits, and logits corresponding to even percents (take sign from head of column used for %)

+	anglit	normit	logit*	−	+	anglit	normit	logit	−
50%	0.0000	0.0000	0.0000	50%	75%	0.5236	0.6745	1.0986	25%
51	0.0200	0.0251	0.0400	49	76	0.5468	0.7063	1.1527	24
52	0.0400	0.0502	0.0800	48	77	0.5704	0.7388	1.2083	23
53	0.0600	0.0753	0.1201	47	78	0.5944	0.7722	1.2657	22
54	0.0801	0.1004	0.1603	46	79	0.6187	0.8064	1.3249	21
55%	0.1002	0.1257	0.2007	45%	80%	0.6435	0.8416	1.3863	20%
56	0.1203	0.1510	0.2412	44	81	0.6687	0.8779	1.4500	19
57	0.1404	0.1764	0.2818	43	82	0.6945	0.9154	1.5164	18
58	0.1606	0.2019	0.3228	42	83	0.7208	0.9542	1.5856	17
59	0.1810	0.2275	0.3640	41	84	0.7478	0.9945	1.6582	16
60%	0.2014	0.2533	0.4055	40%	85%	0.7754	1.0364	1.7346	15%
61	0.2218	0.2793	0.4473	39	86	0.8030	1.0803	1.8153	14
62	0.2424	0.3055	0.4896	38	87	0.8331	1.1264	1.9010	13
63	0.2630	0.3319	0.5322	37	88	0.8633	1.1750	1.9924	12
64	0.2838	0.3585	0.5754	36	89	0.8947	1.2265	2.0907	11
65%	0.3047	0.3853	0.6190	35	90%	0.9273	1.2816	2.1972	10%
66	0.3258	0.4125	0.6633	34	91	0.9614	1.3408	2.3136	9
67	0.3469	0.4399	0.7082	33	92	0.9973	1.4051	2.4424	8
68	0.3683	0.4677	0.7538	32	93	1.0353	1.4758	2.5867	7
69	0.3898	0.4959	0.8001	31	94	1.0759	1.5548	2.7415	6
70%	0.4115	0.5244	0.8473	30	95%	1.1198	1.6449	2.9444	5%
71	0.4334	0.5534	0.8954	29	96	1.1681	1.7507	3.1780	4
72	0.4556	0.5828	0.9445	28	97	1.2226	1.8808	3.4761	3
73	0.4780	0.6128	0.9946	27	98	1.2870	2.0537	3.8918	2
74	0.5006	0.6433	1.0460	26	99	1.3705	2.3263	4.5951	1
75%	0.5236	0.6745	1.0986	25	100%	1.5708	∞	∞	0%

* For more detailed tables see 1953 JASA pp. 568-569.

Note: The difference in overall size is unimportant. (Using, for example, anglit, 4/5 normit, and 1/2 logit, which all *start* out alike, would lead to similar analyses.) What is important is the difference in stretching of the "tails".

We have found it convenient to tabulate an adjusted normit for which

$$z = \text{adjusted normit of } p = \left[\sqrt{\frac{2}{\pi}}\right] \cdot (\text{normit of } p)$$

corresponding to

$$p = \frac{1}{2} \int_{-\infty}^{z} e^{-u^2/\pi}\, du\,.$$

Those concerned with psychological tests and measurement sometimes use a t-score (cp. e.g., McCall 1939, pp. 505-508) which is another expression of a counted fraction belonging to this same mode.

The principal properties of this mode are:

(1) It has no "ends." Positive or negative values of arbitrarily great magnitude correspond to percentages arbitrarily close to 100% or 0%, respectively.

(2) In its effects it lies between the other two modes.

(3) Because of its relation to the famous normal distribution, very extensive tables are available, both of the relation of this mode to others, and of auxiliary quantities, as are many specialized statistical techniques.

(4) For extreme percentages the normit is rather crudely proportional to the square-root of the logarithm of the smaller of the two percentages (the % in the cell considered, or the % *not in* that cell). This relation is not very useful.

Although normits, and probits, have been very popular in other connections, I would surmise, if forced to commit myself, that the mode they represent will not prove to be the most useful of these three modes in most behavioral sciences applications.

The tail-stretchingest of the three modes is that represented by logits, where a logit is defined as the natural logarithm of the ratio of the observed fraction to its complement.

$$\textit{logit} = \log_e \frac{\text{percentage}}{100-\text{percentage}} = \log_e \frac{\text{number of one kind}}{\text{number of other kind}}$$

$$= \log_e \text{ ("odds for")}$$

We found it convenient to tabulate the "half-logit," whose values are just half as large.

The principal properties of this mode are:

(1) It has no "ends." Positive and negative values of arbitrarily great magnitude correspond to percentages arbitrarily close to 100% or 0%, respectively.

(2) For extreme percentages, the magnitude of the logit is quite closely proportional to the logarithm (to any base) of the smaller fraction (or, equivalently, of the smaller percentage).

(3) This mode has certain rather esoteric properties as far as certain types of further analysis are concerned. (These properties, mainly associated with sufficient statistics, are of much greater importance to statisticians developing new analytic techniques than to those who are analyzing actual data. The latter can, of course, make good use of the techniques developed by the former.)

(4) The approximate variance of the logit in simple random binomial sampling can be expressed as

$$\frac{\text{(total number of instances)}}{\text{(number of instances of one kind) (number of instances of the other kind)}}$$

(5) The approximate variance of the half-logit is one-fourth as large, and may be expressed as

$$\frac{\text{(total number of instances)}}{\text{(2 × number of one kind) (2 × number of instances of the other kind)}}$$

(Note that when each class is half the whole, all three numbers in parentheses in this last expression are equal and the approximate variance is just the reciprocal of the total number of instances.)

(6) Fisher's z-transformation of the correlation coefficient r is exactly the logit corresponding to the fraction given by $(1+r)/2$.

There are other interesting properties of logits, but these are at most of modest intuitive significance in the present context. One of these is the relation of logits to the logistic curve, which is the simplest theoretical form for population growth under circumstances where the available resources set an upper limit on the population. (This relation is entirely analogous to that between normits and the cumulative normal distribution.)

U6. COVARIANCES FOR NONINTERSECTING SPLITS

Our examples have included many instances where there were at least two nonintersecting splittings of a group (splittings that together define 3 subgroups rather than 4).

The covariances, or the correlations, between the two fractions, however expressed, are sometimes of interest.

Table 39

Critical table for squared correlation between expressions of two nonintersecting splits of the same simple random sample.

Square of Correlation	Difference in half-logits	Difference in logits
1.00		
	.051	.025
.95		
	.16	.078
.90		
	.27	.13
.85		
	.38	.19
.80		
	.51	.25
.75		
	.64	.32
.70		
	.78	.39
.65		
	.94	.47
.60		
	1.1	.55
.55		
	1.3	.64
.50		
	1.5	.74
.45		
	1.7	.86
.40		
	2.0	.98
.35		
	2.2	1.1
.30		
	2.6	1.3
.25		
	3.0	1.5
.20		
	3.5	1.7
.15		
	4.2	2.1
.10		
	5.2	2.6
.05		
	7.4	3.7
.00		

The situation can be represented as follows (simple random trinomial sampling):

Population fractions	Observed numbers
A	x
B	y
C	z
$A+B+C = 1$	$x+y+z = n$

If we repeatedly draw samples of n observations, n fixed, at random from an infinite population with fractions A, B and C, the observed numbers x, y and z will not be independent. In particular, x and z will be (negatively) correlated. This correlation will be precisely determined by the *difference* between the logits corresponding to the two separations, one into A vs. $B+C$ and one into $A+B$ vs. C.

As a result the two fractions

$$\frac{x}{x+y+z}$$

and

$$\frac{x+y}{x+y+z}$$

will be positively correlated. To a reasonable approximation (especially in large samples) the correlation between the fractions, between the corresponding anglits, between the corresponding normits, or between the corresponding logits will all be the same. This correlation is always positive, its natural logarithm is negative, and the magnitude of this logarithm is the difference between the half-logit for the one split and the half-logit for the other split. A critical table for the square of this correlation is given in Table 39.

V. MODES OF EXPRESSING OTHER QUANTITIES

We have given considerable attention to modes of expression of counted fractions (fractions, %'s, logits, and the like) in Chapter E and Appendix U. It is only fair that we now give a little attention to the results of

experience with various modes of expression for other sorts of values (absolute numbers, amounts, signed amounts, etc.).

V1. EXPRESSING COUNTS

The primeval mode of expression of a count is that represented by the raw count itself. It is clear to almost all of us that the difference between observing 12 instances and observing 13 instances is not as "large" (in the sense of "not as important" or "not as meaningful") as the difference between observing 1 instance and 2 instances. There is a place for modes which compress higher counts together, as compared with lower counts.

The simple square root of the observed number is a representative of a mode of expression which has proved very satisfactory in many circumstances.

Sometimes a more rigorous compression is needed. (This is the case with the number of mites per rat, but not with the number of fleas per rat. Cp. A2 above.) The mode of expression represented by the logarithm (to any handy base) of the count is useful so long as counts of zero are absent. The family of modes of expression represented by

$$\log(\text{count} + \text{constant})$$

where there is a slightly different mode for each positive value chosen as the constant, and where the choice of base of logarithms is unimportant, provides further reasonable alternatives.

These suggestions are empirically useful in quite different situations. They often work. Theoretical support is not necessary.

Sometimes the counts to be dealt with follow a so-called Poisson distribution, to either a close or rough approximation. (The situation may or may not involve additional variability beyond that corresponding to a Poisson distribution.) Both theoretical and empirical justification exists for various modes of expression in such situations. (The behavior of the modes suggested below is much more alike than their appearances suggest.) Most noteworthy are the modes represented by:

(1) log(*count* + *constant*), with a constant close to the average count;

(2) square root of count;

(3) sum of the square roots of (a) the count and (b) the count increased by unity;

(4) a convenient tabulated modification of the last previous (for which see Tukey NYC1).

Any of these is likely to be quite effective, though careful selection among the four in some, rather infrequent, instances, may be worthwhile.

The averages and variances of (4) that apply when the count precisely follows the Poisson distribution may be found in Tukey NYC1 as can a discussion of modes of expression useful in comparing observed and anticipated numbers.

V2. EXPRESSING NON-NEGATIVE AMOUNTS

The possible numerical values of an observation are frequently limited in one direction, but not in the other. By changing the sign of all observed values, if necessary, we can arrange for the limitation to lie in the negative direction. By adding a constant to all observed or modified values, if necessary, we can arrange to have the limitation require precisely that all values be non-negative. When we discuss the natural and convenient modes of expressing non-negative values, we thus cover most, if not all, situations where values are limited on one side.

The values of physical quantities, such as weight, length, and duration, are naturally limited to the left at zero. We can think of such quantities, even though they may be far from typical examples, as paradigms for many more quantities limited to one side but not to the other.

The modes of expression corresponding to the so-called simple family of transformations (Tukey 1957) have proved flexible and useful. Together with all modes represented by

$$(\text{amount} + \text{constant})^{\text{exponent}}$$

the simple family includes such limiting forms as

$$\log(\text{amount} + \text{constant})$$

which fits into the family as if it were the case where the exponent were equal to zero, and

$$e^{-(\text{constant}^*) \cdot (\text{amount})}$$

which fits in as the limiting case where

$$\frac{\text{exponent}}{\text{constant}} = \text{constant}^{\bullet}$$

with both "exponent" and "constant" becoming arbitrarily large.

The natural ladder of modes of expression often descends as follows:

$$\text{amount}$$

$$\sqrt{\text{amount}}$$

$$\log \text{amount}$$

$$\frac{1}{\sqrt{\text{amount}}}$$

$$\frac{1}{\text{amount}}$$

$$\cdots\cdots$$

with each successive step appearing to be of about the same size.

For further discussion, see Tukey 1957 and Tukey NYC1.

V3. EXPRESSING UNRESTRICTED AMOUNTS

The case where the observed values can be both arbitrarily negative and arbitrarily positive has not been studied in any detail, in part because the combination of both-way unlimited values and a need for a different mode of expression does not seem to occur frequently. If a symmetrical mode is desired, those using hyperbolic functions, namely

$$\sinh [(\text{constant}) \cdot (\text{amount})]$$

and

$$\tanh[(\text{constant}) \cdot (\text{amount})]$$

seem to be plausible candidates.

V4. EXPRESSING AMOUNTS RESTRICTED FROM BOTH SIDES

Multiplication of all values by one constant followed by addition of another constant to the result will reduce the general case of amounts restricted to lie between two values to the special case of amounts restricted to lie between 0 and 1, that is, to the special case of fractions.

Alongside any fraction it is natural to consider the complementary fraction (= one minus the first fraction). Symmetry of behavior of fraction and complementary fraction is not guaranteed. But it occurs with reasonable frequency. In a symmetrical situation, any data analyst (probably guided by the simple family of modes of expression for amounts) who thinks of using

$$(\text{fraction})^{\text{exponent}}$$

is compelled by symmetry to give equal attention to

$$(1-\text{fraction})^{\text{exponent}}$$

and finds this most easily done by combining both of these and considering

$$(\text{fraction})^{\text{exponent}} - (1-\text{fraction})^{\text{exponent}}$$

This is quite useful for symmetrical situations, and can be generalized in several ways: (i) by inserting a plus sign followed by a constant inside each parenthesis, (ii) by making the two exponents unequal or the two constants unequal, or by doing both of these, (iii) by inserting a multiplicative constant into either term. All such modes are natural generalizations of the modes of the simple family which we saw to be appropriate for values limited on one side.

V5. RELATION OF MODES FOR RELATIVE NUMBERS TO THOSE JUST DISCUSSED

The usual discussion of anglits, normits, and logits (cp. E4) is likely to spend some attention on their behavior for extreme fractions. It is easy mathematics to derive limiting forms in which extreme anglits are proportional to $\sqrt{p}$, extreme normits to $\sqrt{\log p}$ and extreme logits to

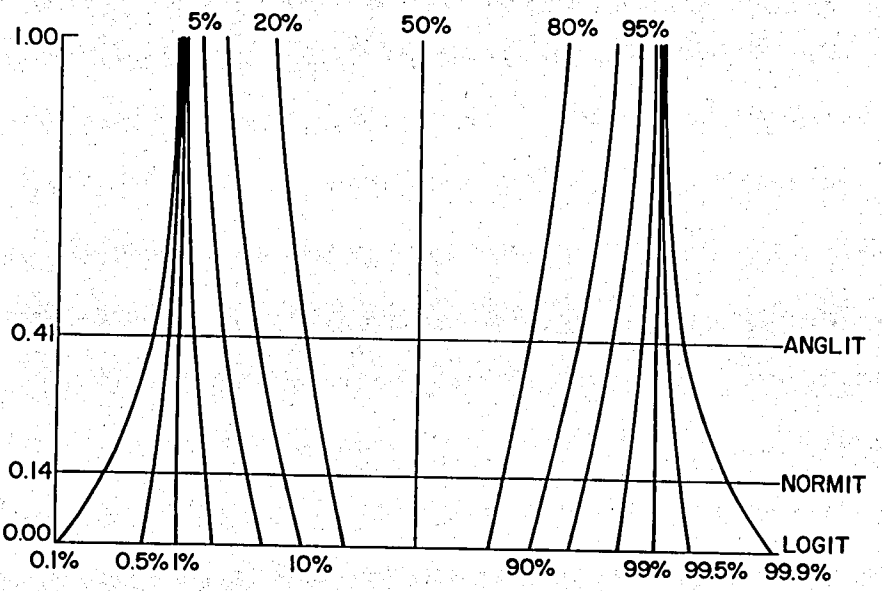


Figure 16. Comparison of modes of expression for fractions based on counts with the very simple symmetric modes for general fractions

log p. These results are of little use in practical data analysis. Observed fractions below 1% or above 99% are rarely determined from counts with sufficient accuracy to make such asymptotic results useful.

Far more guidance can be obtained by relating anglits, normits and logits to the very simple symmetrical modes of the last section. Empirically we have the approximations

$$\begin{aligned} \text{anglit } p &\sim [p^{.41} - (1-p)^{.41}] \\ \text{normit } p &\sim [p^{.14} - (1-p)^{.14}] \\ \text{logit } p &= \log p - \log(1-p) \end{aligned}$$

which hold over the range $0.01 \leqslant p \leqslant 0.99$ with surprising accuracy. (Cp. Tukey 1960)

Figure 16 shows the behavior of

$$\text{constant}[p^{\text{exponent}} - (1 - p)^{\text{exponent}}]$$

for exponents of 1, 1/2, 0.41, 0.14 0.00 (i.e. logarithm), and $-1/2$ in comparison with suitable multiples of anglits, normits, and logits. (All multiplicative constants have been adjusted to bring 1% and 99% to the same two values.)

W. INSTRUMENTALITY AND CAUSALITY

The point that empirical evidence alone cannot establish causality was made briefly but firmly in Al. This appendix gives additional support to the general argument by analyzing in some detail a particular attempt to use empirical evidence to establish causality. This seemed worthwhile because this particular attempt is relatively novel and may tend to attract appreciable attention during the next few years.

A certain amount of structure has to be discussed as a necessary preliminary. Fortunately this structure has considerable interest for its own sake.

W1. REGRESSION

One of the most classical and most powerful techniques of statistics is regression. (Yet we must agree with Cochran that it is probably the most poorly taught and expounded.) If we have a sample, or even a population, of pairs of associated numerical values (x, y), which may be, for example, father's height and son's height, or price and sales, or rainfall and crop yield, it is natural to ask how to predict the one from the other. Given x, about what value can we anticipate for y? Given y, about what value can we anticipate for x? These are the classical questions; others are easily added.

In the simplest situations it is satisfactory to "predict" y from x using a linear relation

$$y = a + bx \ .$$

Satisfactoriness, of course, does not mean that y will be correctly predicted in every instance. Rather it means that no other function of x

will do much better. And this means that if we collect all (x, y) pairs with a given value of x, and examine the distribution of their y-values, then $ax + b$ gives a useful typical value for this distribution, or at least about as useful a typical value as we know how to compute from our limited body of data.

Regression problems are not confined to situations where everything is normally distributed. Far from it. But almost everything is simplest in such situations which are also sufficiently general to illustrate our points. So we shall be unrealistic and confine our discussion to cases with much normality.

If the (x, y) pairs follow a bivariate normal distribution, then the y's for fixed x follow a univariate normal distribution. Since this distribution is symmetrical its natural typical value is its center, which is both median and average. Since the distribution is normal its description is completed by giving its average and its variance. Thus the distributions of y for given x are completely described by a combination of normality of shape with two functions of x, the average and the variance. It turns out that bivariate normality for (x, y) not only implies normality for y given x, but it also implies constant variance of y given x, and linear dependence of "average y given x" upon x.

Since "bivariate normality for (x, y)" is symmetrical in x and y, the same must hold with y and x interchanged. Thus the two simple regressions

$$\text{ave}\{y \text{ given } x\} = a + bx$$
$$\text{ave}\{x \text{ given } y\} = a' + b'y$$

are exactly linear. If we plot observed values of (x, y) and the two corresponding regression lines, we obtain a picture such as that of Figure 17 where the two regression lines do not coincide. When such a picture is first seen, it is natural to blame this apparent lack of agreement on something which might be altered, perhaps inadequate size of sample, perhaps inadequate theory. But such an attitude is quite wrong; the disagreement of the simple regression lines is an essential feature of regression. We can see that this is so by looking at an extreme case.

Suppose x and y are statistically unrelated. To be more specific let them have normal distributions with averages ave $x = 3.1$, ave $y = 5.3$ and variances var $x = 1.50$ and var $y = 0.79$. (Their covariance will of course be zero.) Since x and y are independent, giving x does not

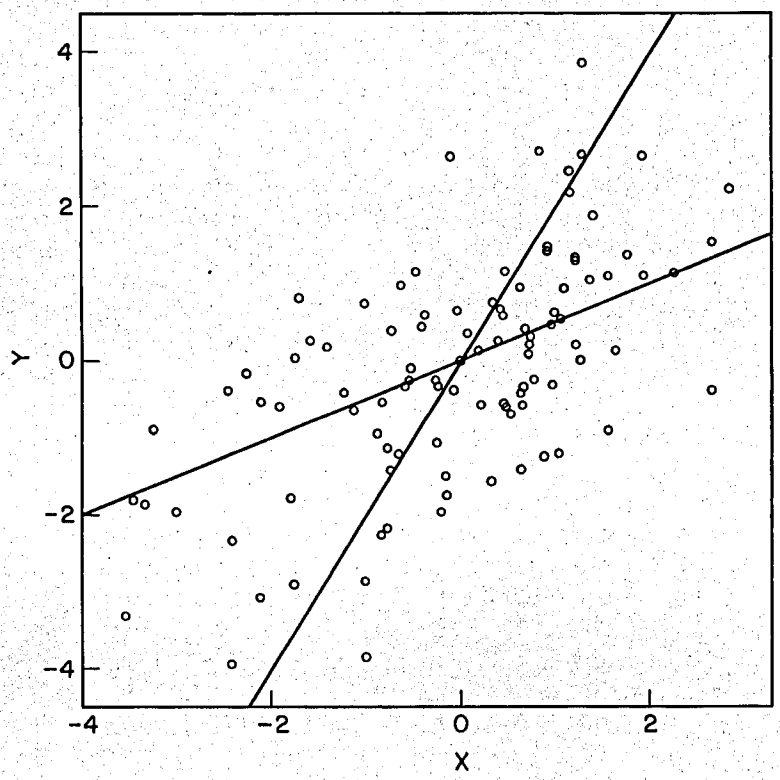


Figure 17. Plot of 100 pairs (x, y) showing the two regression lines

affect the average of y, and vice versa, so that the simple regression lines are

$$\text{ave}\{y \text{ given } x\} = 5.3,$$
$$\text{ave}\{x \text{ given } y\} = 3.1$$

One regression line is horizontal, the other vertical. Yet both are conveying the same message "the value of the one quantity tells you nothing about the (average) value of the other."

W2. "ERRORS" AND STRUCTURAL VARIATES

Suppose now that y is "measured with error." That is, what is measured is v, where

$$v = y + e$$

where "e" stands for an error or fluctuation. In the simplest case, which will suffice to make our points, e is normally distributed, is in fact normally distributed with average 0 and some fixed variance, both unconditionally and given either or both of x and y. Since the average of e given x is zero, the average of v given x is the same as the average of y given x. Consequently, the regression of the measurement v on x is identical with the regression of the concealed quantity y on x. A similar result does *not* hold for the regression of x on v.

Introduction of error or fluctuation into the measurement of y attenuates (weakens) the regression of x on y. If

$$\begin{aligned} \text{ave}\{x \text{ given } y\} &= a' + b'y \\ \text{ave}\{x \text{ given } v\} &= a'' + b''v \end{aligned}$$

then b'' falls between 0 and b'. More precisely

$$b'' = \frac{\text{var } y}{\text{var } y + \text{var } e} b' = \frac{\text{var } y}{\text{var } v} b' .$$

In many instances, errors or fluctuations can occur in the measurement of either or both of x and y. In general we should put

$$\begin{aligned} u &= x + e' \\ v &= y + e \end{aligned}$$

where e' and e both represent errors or fluctuations, and may hopefully be assumed independent of *both* x and y. There are situations where e and e' are statistically related. These are usually more difficult situations. We shall be wise to restrict our attention here to the case where e and e' are statistically independent.

We saw that the variability of e left the regression of v on x the same as that of y on x, while that of x on v was attenuated with respect to that of x on y. By symmetry the regression of u on v is the same as that of x on v. Thus the regression of u on v is attenuated from that of

x on y, as it is easy to show, by an amount depending specifically on the variance of e. Similarly the regression of v on u is attenuated from that of y on x by amount depending specifically on the variance of e'.

In such a situation it is reasonable to call u and v the ***measurable variables,*** and x and y the ***structure variables.*** It is easy to estimate the simple regressions of each measurable variable on the other, if we have a proper supply of observed pairs of values (u, v). It may be of interest to estimate the structural regressions, the simple regressions of each structural variable on the other. This cannot be done from a simple collection of (u, v) pairs alone. Something more must be added. The simplest addition, one not too likely to be available, is adequately precise knowledge of both error variances, the variance of e' and the variance of e. For if these be known, the amount of the attenuations can be calculated, and corrected for. If, in particular, one error is absent, one structural regression will coincide with the corresponding measurable regression.

W3. WHY MAY STRUCTURAL RELATIONS BE INTERESTING?

It is a characteristic of the scientific approach, whether this approach be physical, biological, or behavioral, to seek, wherever possible, an understanding of mechanisms, of underlying factors rather than surface appearances. In many economic situations, for example, it is feasible to get overt information such as prices and volumes and to seek to penetrate to the underlying economic mechanisms. But little thought is required to see that information about the operation of simple economic mechanisms is bound to be obscured in the measurable variables. Clerical errors and discrepancies among definitions (Morgenstern, 1950) undoubtedly provide the final wave of concealing fog. But the inevitable differences between the simple mechanisms with which we have to begin the study of any situation and the complex mechanisms of the true situation are not likely to be negligible. Some of these differences will be behavioral, perhaps involving group phenomena, perhaps involving the superstitions of someone with a great personal effect on the market. Others may be biological, like an epidemic of influenza, or physical, like a widespread fog. All contribute to differences between the hypothetical and simply-behaving structural variables and the actual and complexly-behaving measurable variables.

It is true that we will wish, in due course, to advance from an understanding of the simple approximate mechanisms toward an understanding of the complex actual mechanisms. But we must begin if

we are to proceed. We must get hold of the simple approximate mechanisms first.

Almost all economic observations are nonexperimental; this is one foundation for the importance of assessing structural regressions in economics. If we could reach in and change the structural variables directly, we could assess the structural regressions in much simpler and easier ways.

Many of the results of economics are intended to apply to structural change; this is a second foundation for the importance of assessing structural regressions in economics. Whether the attempt is to infer behavior in quite a different market, or in one differing in commodity, in epoch, or in economic system, or whether the attempt is to infer the results of specific changes in economic policies or practice, the difference between the situation in which the data were gathered and the situation to which the conclusions hopefully apply is likely to be at least a structural one.

Measurable regressions are appropriate for predictions to be used under exactly the same circumstances as they were obtained. Structural regressions are appropriate for predicting the effect of changed circumstances.

W4. *INSTRUMENTAL VARIATES AND INSTRUMENTAL CLASSIFICATIONS*

The most effective methods of assessing structural regressions revolve around the notion of an instrumental variate, which has recently been expanded to the notion of an instrumental classification. The notion is simple; the manner in which it succeeds is more subtle; the conditions which must hold for it to operate properly are most subtle of all.

Suppose that the structural variables x and y are not alone, that there is also a structural variable z which is related (at least statistically) to x, *and* to y. (It is quite possible for z to be identical with, or precisely determined by, either x or y.) And suppose further that there is an observed variable w which differs from x by errors and fluctuations:

$$w = z + e''$$

and, *most vitally,* that these errors and fluctuations are such that *both* e *and* e' are statistically *independent* of *both* z *and* e''. Thus w can be used

to tell something about x, something about y, but nothing about either e' or e. This is the simple concept.

Now let the values of w be quantitative. (These may perfectly well be simple quantitative scores applied to the classes of an ordered classification.) It is then appropriate to consider regressions on w, in particular the regressions of v on w and of u on w. These will be attenuated forms of the regressions of y on z and of x on z. In both instances, the attenuation will come from the variance of e''. It is easy to show that both attenuations are by the same ratio. Consequently, under the basic independence-of-error assumption,

$$\frac{\text{slope of } y \text{ on } z}{\text{slope of } x \text{ on } z} = \frac{\text{slope } v \text{ on } w}{\text{slope } u \text{ on } w}$$

so that the left-hand ratio of slopes can be estimated.

If now the dependence of y on z can be considered as all passing through x, we will have attained one of our goals. Indeed, a necessary condition for that sort of dependence is

$$\text{slope of } y \text{ on } x = \frac{\text{slope of } y \text{ on } z}{\text{slope of } x \text{ on } z}$$

so that one of the structural regression coefficients, the slope of y on x, can be estimated by the ratio of any estimates of two of the measurable regression coefficients, the slopes of y and x on z.

When will it be reasonable to assume that the dependence of y on z can be considered to pass through x? If y were exactly and linearly determined by x, this would be the case. If z were identical with x, or, more generally, if x were exactly and linearly determined by z, this would also be the case. These situations appear special. But they are exactly the situations which may appear in a wide variety of problems. Thus, in a supply-demand situation, the simplest approximate model relates demand to price exactly. This amounts to making y an exact function of x. Even more frequent is the case where z is equal to, or a function of x, where w and v are measures of the same thing, one being perhaps a much cruder measure than the other.

This last situation brings us to the most subtle point; why does the technique ever work in practice? For we may take $z = x$, yet we dare not take $w = u$. And w can be a quite crude measure of x, but we dare not let w equal u plus additional fluctuations and errors. The root of the matter lies in the assumption of independence of errors, in the failure of z, e'' or w to reveal anything at all about e' or e. This

assumption is not to be proved by empirical observation or analysis of data. It can only be a result of insight or theoretical argument.

Indeed it is usually true that when instrumental variates are used to obtain structural regressions, the instrumental variate serves to define the structural variables themselves, to define them as those variables for which "observed variable minus structural variable," now to be called "error," is independent of instrumental variable. If the instrumental, variables are chosen wisely and carefully, and not in desperation, are chosen to estimate what is really desired, and not anything at all except the simple regressions, then their role in defining the structural variable is usually helpful and desirable, rather than dangerous or unpleasant.

W5. *INSTRUMENTAL CLASSIFICATIONS AND THE WORKING-WOLD ANALYSIS*

Some instrumental variables are not quantitative; quantitative instrumental variables are sometimes used qualitatively. In either case we speak of an instrumental classification. In either case the formula $w = z + e''$ and the distinction between w and z become irrelevant at best. And the independence-of-error assumption reduces to the independence of both e' and e from w. To grasp the structural regression we still need to assume that the whole of the connection between w and y passes through x.

Sorting out observed pairs according to values of w can serve some of the purposes of experimentation. The basic argument is simple. The classifying instrumental variable is independent of e' and e. Thus if we sort out observed pairs because of the corresponding values of w, and for no other reason, we shall have sorted in terms of the values of x and y alone. This is a way of reaching in and grasping the structural variables. If the relationships of x and y to z are weak, the grasp may be feeble, but it is there, and it allows us to do some of the things which would be easy if experimentation were possible.

It is now relatively easy to inquire whether either structural regression is identical with the corresponding measurable regression. The slope of y on x will be the same as the slope of v on u if and only if the error e' in x vanishes. If there is no error e' in x for the whole population, there is no error in x for any subpopulation. In particular, there is none for any subpopulation sorted out in terms of the values of w. If we sort out several subpopulations and find the slope of v on u substantially the same within each, this is evidence, ***not conclusive but well worth attention,*** that the structural regression of y on x is the same as the simple regression of v on u. If, on the contrary, we find the

slope of v on u changing substantially from one sorted-out subpopulation to another, we conclude that the structural regression is definitely different (and hence stronger) than the simple regression of v on u.

What subpopulations shall we sort out, and how shall we examine the corresponding slopes? At least two choices are worth description.

One approach was pioneered by Holbrook Working (Working 1933, Working 1934) and has recently been taken up by Wold (Wold 1961). In this approach a number of nonoverlapping, moderately small subpopulations are sorted out, the slope of v on u estimated for each, and the result summarized by a measure of variability of these slopes, perhaps the ratio of the standard deviation (between sorted-out subpopulations) of the estimated slopes to their mean.

The attenuation due to error in x is by the factor

$$\frac{\text{(mean square deviation of } x\text{'s)}}{\text{(mean square deviation of } x\text{'s)} + \text{(mean square deviation of errors)}}$$

which depends only upon

$$\frac{\text{(mean square deviation of } x\text{'s)}}{\text{(mean square deviation of errors)}}$$

and it is upon changes in this latter ratio that the whole instrumental classification rests. If the assumed independence of error from w holds completely, then the mean square deviation of errors will not depend upon w and the denominator of the last ratio will be the same for every sorted-out subpopulation. In such circumstances, if the sorted-out subpopulations all have very similar values of the mean square deviation of x, the ratios will be nearly the same for all sorted-out subpopulations, as will the slopes. Thus the Working-Wold approach will not be effective if errors are truly independent of the sorting variable, *and* the sorted-out subpopulations have similar dispersions for x, or for u.

There is, however, some compensation for this weakness. For effective use of the instrumental device, it is only necessary that w tell us nothing about e' *in a linear way*. It is sufficient to require absence of (Pearsonian) correlation in place of absence of dependence. And this can hold while the mean square deviation of e' in the various sorted-out subpopulations vary among themselves. If now the mean square deviations of x in the sorted-out subpopulations are all about the same, the ratios will differ, and evidence that the simple regression is not the structural regression can be gathered.

W6. *ANOTHER APPROACH, AND THE ANALYSIS OF RECIPROCAL SLOPE*

When the mean square deviation within sorted-out subpopulations is constant, in particular when e' is completely, and not merely linearly, statistically independent of w, another approach is possible, and seems likely to be more sensitive. The actual slope of v on u will be proportional to the factor by which the slope of y on x has been attenuated. Passing to the reciprocals:

$$\frac{1}{\text{slope of } v \text{ on } u} \propto \frac{MSD\{x\} + MSD\{e'\}}{MSD\{x\}}$$
$$\propto 1 + MSD\{e\} \cdot \frac{1}{MSD\{x\}}$$

where "MSD" stands for the mean square deviation.

If we plot

$$\frac{1}{\text{slope of } v \text{ on } u}$$

against

$$\frac{1}{\text{mean square deviation of } x\text{'s}}$$

for various sorted-out subpopulations, we should expect to find a roughly linear relationship, and should be able to take the ratio of slope to vertical intercept of this line as an estimate of the mean square deviation of the e's. To make this process effective, we wish to choose the sorted-out subpopulations so that:

(1) the estimates of slope of v on u are stable;

(2) the mean square deviation of x varies substantially, or as is often equivalent, the mean square deviation of u varies substantially.

To do this we will be wise to accept overlapping subpopulations of different sizes, and to consider seriously making a number of dissections of the population into subpopulations, with the sizes of the mean square deviation for u about the same for the subpopulations of an individual dissection, but quite different from dissection to dissection. More detailed discussion of this approach should await further trial on actual data.

It would not be right to close this discussion, however, without pointing out that many instrumental approaches can be regarded as various instances of looking to see if the sum of squares of x follows the sum of cross-products of x and y quantitatively, just as the little lamb followed Mary qualitatively.

W7. STRUCTURAL INFORMATION IS NOT CAUSAL INFORMATION

We have seen a little of the possibilities and difficulties of the search for information about structural regressions; we must deal now with its greatest temptation. It is at least interesting, and sometimes very important, to know which structural variable has been measured with error or fluctuation, and which, if one there be, has been measured cleanly. This information can sometimes be combined with theoretical insight into the subject-matter to throw light on questions of causality. The danger and the temptation is to try to establish ***causality*** by using such empirical evidence alone.

In this section we try to reveal this danger by example. Consider first a farm on which there are a number of piles of stones of varying weight. The weights vary within each pile. And the average weight of a stone varies substantially (i.e., more than would correspond to random choice) from pile to pile. Strangely enough, each stone is clearly and ineffaceably labeled with a serial number. An investigator of stone weights is coming to weigh these stones. He is so careful that he insists in weighing each stone twice, once on each of two weighing machines. And in order to keep subjective errors to a minimum he insists on weighing all the stones on one weighing machine before beginning to weigh on the other.

Suppose further that one weighing machine is very precise, while the other is subject to considerable random error. The change in true weight of any stone between the two weighings is negligible. Both structural variables are the same, the true weight of the stone. One observed variable, the observed weight with the precise weighing machine, is very closely the same as the corresponding structural variable. The other observed variable, measured with the imprecise weighing machine, is not closely the same as the corresponding structural variable. Any sensible analysis of the data, for example, one using "pile" as an instrumental classification, or one using "mean weight of other stones in the same pile for both weighings" as an instrumental variate, will discover that the measurable regression of imprecise weighing on precise weighing may possibly be a structural

regression but that the measurable regression of precise weighing on imprecise weighing cannot be structural.

Such a result gives no information at all about causality. If the investigator uses the precise weighing machine first, the first weighing will appear to give the structural answer. If, on the other hand, he uses the imprecise one first, the second weighing will give the structural answer. But the order of use of weighing machines can have nothing at all to do with the direction of causality, with the nature of the causal relations among weights.

If causality makes any sense here, the earlier weight is surely the cause of the later weight. And we have seen that either structural result can be obtained.

This example may perhaps be objected to because it is felt that causality is not a valid concept in this particular situation. It is easy to modify it slightly to avoid this difficulty, but it may be more helpful and illuminating to modify it substantially. Let us replace stones by men and the earliest born of their male children to reach the age of 18. We can replace the piles of stones by groups of men living in ethnically different parts of the world. We suppose the male parents to be weighed when they are 18, while the male children are weighed some decades later, when they are themselves 18. Again, one measurement is made with an imprecise weighing machine and one with a precise one. Again, whichever measurement, of parents, or of children, is made with the precise weighing machine will turn out to be possibly structural. Here the direction of causality is unequivocal; the weights of children to come cannot cause the weight of the father at age 18.

Structural information cannot be directly converted into causal information.

X. REFINING ADJUSTMENT FOR BROAD CATEGORIES

In Section D3 we showed how the use of broad classes may not suffice to *eliminate* the effect of some variable, although it is usually very helpful. It is natural to ask how we can do better. This appendix attempts to provide one way to do better, a way which is novel, and which will require considerable trial before we can be sure of its efficacy, but one whose apparent efficacy is considerable.

X1. THE APPROACH

As in so many approaches to a new technique, we are going to proceed *as if* there were an underlying quantitative variable which has a normal distribution. Notice that ***it has not been said*** that we ***assume*** the existence of a normally-distributed underlying quantitative variable. It is important that we have not said this. To say it would be to take a narrow, purely mathematical approach to a broad problem whose essentials are not mathematical. (But in whose solution we look to mathematics for aid.)

A physical example may illuminate the situation. How does one begin to treat the motion of the planets around the sun? One begins by treating each object, planet or sun, as if it were a "point-mass." This does not mean that the physicist or astronomer is ***assuming*** that all the mass of the sun is concentrated at a point. Far from it. He is, instead, treating first as simple a case as seems likely to provide the essentials of the answer. To his point-mass solution (if he can solve the problem of *n* bodies!) he has an obligation to add consideration of how well this solution is likely to provide all the essentials. Part of this consideration should come from his own professional understanding of the situation, specifically of what is likely to be how important; another part may come from trial solutions of slightly more complex situations, from a study of perturbations, or from comparison with experiment. But there will be no substitute for a ***combination*** of a solution (of a problem that is, in almost every instance, much simpler than the real situation) ***and*** a consideration of the likelihood of adequate applicability to the real situation of this solution (and not of the hypotheses from which it was derived).

Data analysis is not different from physical science in this respect; procedures of data analysis are usually found be seeking something which is reasonable in a very special case, and then validating it (as much as may be appropriate) by ***both*** professional appraisal of the likelihood of its adequacy as a working approximation and trial in diverse practical circumstances.

The very simple circumstances which are so often our initial concern are not assumptions, but rather guides, guides on trial rather than guides fully accepted. It is most important, not only in this instance, but throughout data analysis, to understand this fact, and to approach the synthesis and appraisal of data-analytical techniques with corresponding attitudes and tools.

X2. THE STRUCTURE

If we are to correct more effectively for the effects of a variable which is known only in terms of broad classes, we must do something better than treating the mean behavior of all instances which fall into a broad class as though they fell at the *center* of the broad class. We must treat the mean behavior as falling at an appropriate point and then allow for the fact that this point is not the mid-class point.

If we know that the interest of 1194 men in a forthcoming election divided 449, 789, and 56 between great, moderate, and none, we can assign percentage positions to the breaks and then, using tables of the standard normal cumulative, we can assign normit values to the breaks. For the broad class between the two breaks ("moderate interest") we can easily determine a mid-class normit as the arithmetic mean of the adjacent break normits. The other two classes are open-ended, and have no mid-class point. We want standard points for these classes also.

We have little choice but to fix these outer standard points at a prescribed distance outside the extreme breaks. It is convenient to choose this distance as one-quarter the mean width of the classes between the extreme breaks.

All this computation goes forward in normits, and is applicable to the specific group treated. Once we can replace actual mean behavior for broad classes by mean behavior adjusted to the standard points, we will be in shape to make comparisons from one group to another. In doing this we need pay no further attention to the normit values, which have served their purpose by allowing us to make adjustment.

Several points deserve stress. First, there is no necessary connection between the normit scales used to adjust the different groups. Second, the adjustments should be small, and are reasonably made by linear interpolation. Third, the whole procedure assumes that it *is* reasonable to think of a single underlying continuous variable, with smooth, singly-humped distribution. (It *might* not be wise to use such an adjustment when the broad classes are, for example, "working class," "middle class," "upper class," since the breaks between these classes might be so well defined that a reasonable underlying continuous variable would have to have dips in its density of distribution near each break.)

The first of these three points is emphasized in Table 40, which illustrates the construction of normit scales and standard points, for reported interest in a forthcoming election, for 1294 men and 1418 women. (Data of Lazarsfeld, Berelson, and Gaudet 1948, as reported by Hyman 1955, page 297, Table 27. This example is also discussed in Section E9.)

Table 40

Construction of normit scales and location of standard points for two groups, one of 1294 men, the other of 1418 women.

Reported interested	Number of cases	Breaks in %	Breaks in normits	Std. points
----------	1294 men	----------		
Great	449			0.93
		65.4%/34.6%	0.40	
Moderate	789			−0.66
		4.3%/95.7%	−1.72	
None	56			−2.25
----------	1418 women	----------		
Great	328			1.16
		76.8%/23.2%	0.74	
Moderate	852			−0.11
		16.8%/83.2%	−0.96	
None	238			−1.38

$-0.66 = \frac{1}{2}[(0.40)+(-1.72)]$, $-0.11 = \frac{1}{2}[(0.74)+(-0.96)]$,

$0.53 = \frac{1}{4}[(0.40)-(-1.72)]$, $0.42 = \frac{1}{4}[(0.74)-(-0.96]$,

$0.93 = 0.40 + 0.53$, $-2.25 = -1.72 - 0.53$,

$1.16 = 0.74 + 0.42$, $-1.38 = -.96 - 0.42$.

X3. THE APPROPRIATE POINTS

To lay out our normit scales, and to establish the corresponding standard points we needed only tables relating deviation to break (to cumulative fraction) for the standard normal distribution (the one with average zero and variance one). Such tables are to be found in almost every statistics book.

To determine appropriate points at which we may think of the cases in a broad class as concentrated, we need, essentially, a table of centers of gravity of segments of the standard unit normal. A table of this latter sort has been given by Leverett (1947). For our present purposes, a modification of Leverett's table, giving the displacement of the appropriate point from the class midpoint (for extreme classes, from

the class boundary) is even simpler to use. Such a table is given as Table 41.

Table 41

Deviation of means of segments of the standard normal distribution from means of class boundaries (from the class boundary when there is but one). (Units determined so that $\sigma = 1.0$.)

Open-ended classes		Deviation inward from mean of class boundaries for interval classes					
	Deviation beyond class boundary	One	The other tail				
Prob.	(outward)	tail	5%	10%	20%	40%	70%
1%	.34	1%	.26	.36	.41	.43	.32
2	.37	2	.15	.24	.31		
3	.39	3	.08	.18	.24	.27	
4	.40	4	.02	.13	.20		
5	.42	5	.00	.10	.17	.18	.11
6%	.43	6%	.03*	.07	.14		
8	.45	8	.07*	.03	.10		
10	.47	10	.10	.00	.07	.09	.04
12	.49	12	.12	.02*	.05		
15	.52	15	.14	.05	.03	.05	.02

*Interpolated as though there were a change in sign.

(to be cont'd and modified)

Applying this table to the example already begun, we find

(1) that the "appropriate point" for the 449 men with "great interest" is 0.67 normit beyond the break, and hence falls at $0.40 + 0.67 = 1.07$ normits, and

(2) that the "appropriate point" for the 789 men with "moderate interest" falls about 0.21 normit from the mid-class point, namely at $-(0.66-0.21) = -0.45$.

Continuing, we find the results shown in Table 42. Table 43 sets out the original and adjusted comparisons.

The effects of adjustment are not large (recall that Table 11 found the standard errors of the differences to be $\pm .31$, $\pm .14$ and $\pm .19$, respectively), but they are not negligible. (The shift in weighted

average, 0.19, is almost twice the standard error, 0.11, of the unadjusted difference.) Interestingly enough, the differences for the three classes of reported interest are more consistent after adjustment.

As noted earlier, this method of adjustment is on trial. One test means little, but this one is at least encouraging.

Table 42

Mean responses, appropriate points, standard points, and adjusted mean responses for 1294 men and 1418 woman.

Reported interest	Response (half-logit voting)	Appropriate point	Standard point	Interpolated response*
		1294 men		
Great	2.30	1.07	0.93	2.27
Moderate	1.95	−0.45	−0.66	1.81
None	0.79	−2.13	−2.25	0.71
		1418 women		
Great	1.95	1.33	1.16	1.83
Moderate	0.95	−0.08	−0.11	0.93
None	−0.12	−1.50	−1.38	−0.03

$$*2.27 = 2.30 - \frac{1.07 - 0.93}{1.07 - (-0.45)}(2.30-1.95)$$

$$1.81 = 1.95 - \frac{-0.66 - (-0.45)}{-2.13 - (-0.45)}(1.95-0.79)$$

$$0.71 = 0.79 - \frac{-2.25 - (-2.13)}{-2.13 - (-0.45)}(1.95-0.79)$$

and so on.

Table 43

Effect of adjustment upon comparison of men and women as to voting fraction. (All values in half-logits.)

Reported interest	Male response	Female response	Difference
----------unadjusted responses----------			
Great	2.30	1.95	0.35
Moderate	1.95	0.95	1.00
None	0.79	−0.12	0.91
(weighted*)			(0.90)
----------adjusted responses----------			
Great	2.27	1.83	0.44
Moderate	1.81	0.93	0.88
None	0.71	−0.03	0.68
(weighted*)			(0.71)

* With weights 2, 10 and 5, respectively, which are sufficiently closely proportional to the reciprocals of estimated variances.