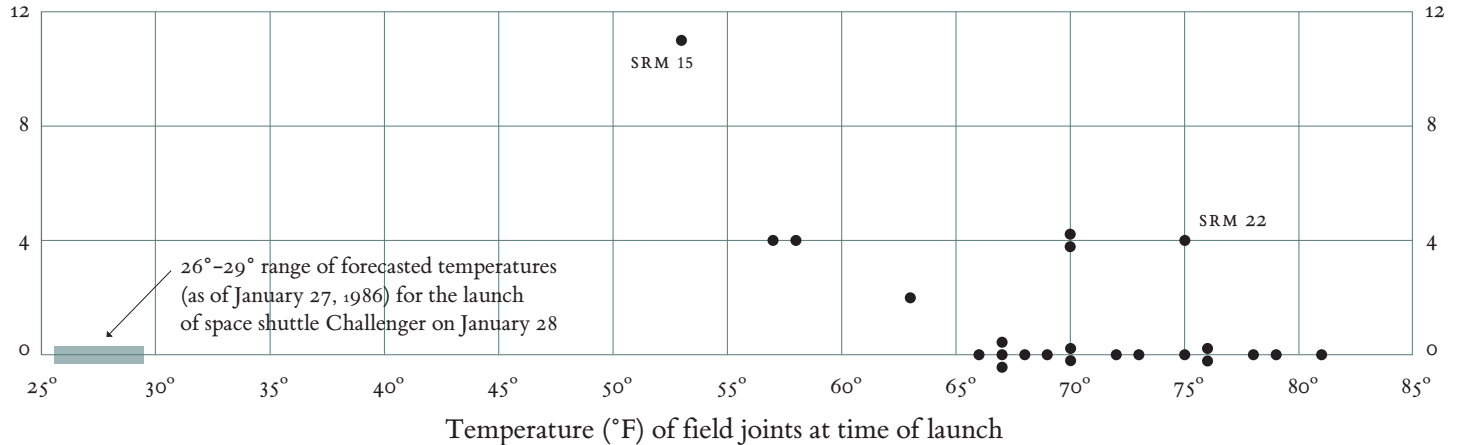


EDWARD R. TUFTE

VISUAL AND STATISTICAL THINKING: DISPLAYS OF EVIDENCE FOR MAKING DECISIONS

O-ring damage
index, each launch



JOHN SNOW AND THE CHOLERA EPIDEMIC

THE DECISION TO LAUNCH THE SPACE SHUTTLE CHALLENGER



Edward Tufte has written eight books, including *Beautiful Evidence*, *Visual Explanations*, *Envisioning Information*, *The Visual Display of Quantitative Information*, *Data Analysis for Politics and Policy*, *Political Control of the Economy*, and *Size and Democracy* (with Robert A. Dahl). He writes, designs, and self-publishes his books on information design, which have received more than 30 awards for content and design. He is Professor Emeritus at Yale University, where he taught courses in statistical evidence, information design, and interface design.

Copyright © 1997 by Edward Rolf Tufte
PUBLISHED BY GRAPHICS PRESS LLC
POST OFFICE BOX 430, CHESHIRE, CONNECTICUT 06410
WWW.TUFTE.COM

All rights to illustrations and text reserved by Edward Rolf Tufte. This work may not be copied, reproduced, or translated in whole or in part without written permission of the publisher, except for brief excerpts in connection with reviews or scholarly analysis. Use in *any* form of information storage and retrieval, electronic adaptation or whatever, computer software, or by similar or dissimilar methods now known or developed in the future is also strictly forbidden without written permission of the publisher.

Introduction

THIS booklet, meant for students of quantitative thinking, reproduces chapter 2 of my book *Visual Explanations: Images and Quantities, Evidence and Narrative*.

The general argument is straightforward:

An essential analytic task in making decisions based on evidence is to understand how things work—mechanism, trade-offs, process and dynamics, cause and effect. That is, intervention-thinking and policy-thinking demand causality-thinking.

Making decisions based on evidence requires the appropriate display of that evidence. Good displays of data help to reveal knowledge relevant to understanding mechanism, process and dynamics, cause and effect. That is, displays of statistical data should directly serve the analytic task at hand.

What is reasonable and obvious in theory may not be implemented in the actual practice of assessing data and making decisions. Here we will see two complex cases of the analysis and display of evidence—the celebrated investigation of a cholera epidemic by Dr. John Snow and the unfortunate decision to launch the space shuttle Challenger.

Edward Tufte

Although we often hear that data speak for themselves, their voices can be soft and sly.

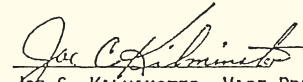
Frederick Mosteller, Stephen E. Fienberg, and Robert E. K. Rourke, *Beginning Statistics with Data Analysis* (Reading, Massachusetts, 1983), 234.

Negligent speech doth not only discredit the person of the Speaker, but it discrediteth the opinion of his reason and judgment; it discrediteth the force and uniformity of the matter, and substance.

Ben Jonson, *Timber: or, Discoveries* (London, 1641), first printed in the Folio of 1640, *The Workes . . .*, p. 122 of the section beginning with *Horace his Art of Poetry*.

MTI ASSESSMENT OF TEMPERATURE CONCERN ON SRM-25 (51L) LAUNCH

- 0 CALCULATIONS SHOW THAT SRM-25 O-RINGS WILL BE 20° COLDER THAN SRM-15 O-RINGS
- 0 TEMPERATURE DATA NOT CONCLUSIVE ON PREDICTING PRIMARY O-RING BLOW-BY
- 0 ENGINEERING ASSESSMENT IS THAT:
 - 0 COLDER O-RINGS WILL HAVE INCREASED EFFECTIVE DUROMETER ("HARDER")
 - 0 "HARDER" O-RINGS WILL TAKE LONGER TO "SEAT"
 - 0 MORE GAS MAY PASS PRIMARY O-RING BEFORE THE PRIMARY SEAL SEATS (RELATIVE TO SRM-15)
 - 0 DEMONSTRATED SEALING THRESHOLD IS 3 TIMES GREATER THAN 0.038" EROSION EXPERIENCED ON SRM-15
 - 0 IF THE PRIMARY SEAL DOES NOT SEAT, THE SECONDARY SEAL WILL SEAT
 - 0 PRESSURE WILL GET TO SECONDARY SEAL BEFORE THE METAL PARTS ROTATE
 - 0 O-RING PRESSURE LEAK CHECK PLACES SECONDARY SEAL IN OUTBOARD POSITION WHICH MINIMIZES SEALING TIME
- 0 MTI RECOMMENDS STS-51L LAUNCH PROCEED ON 28 JANUARY 1986
 - 0 SRM-25 WILL NOT BE SIGNIFICANTLY DIFFERENT FROM SRM-15


JOE C. KILMINSTER, VICE PRESIDENT
SPACE BOOSTER PROGRAMS

MORTON THIOKOL INC.
Wasatch Division

INFORMATION ON THIS PAGE WAS PREPARED TO SUPPORT AN ORAL PRESENTATION
AND CANNOT BE CONSIDERED COMPLETE WITHOUT THE ORAL DISCUSSION

The final approval and rationale for the launch of the space shuttle Challenger, faxed by the rocket-maker to NASA the night before the launch. The rocket blew up 12 hours later as a result of cold temperatures.

Visual and Statistical Thinking: Displays of Evidence for Making Decisions

WHEN we reason about quantitative evidence, certain methods for displaying and analyzing data are better than others. Superior methods are more likely to produce truthful, credible, and precise findings. The difference between an excellent analysis and a faulty one can sometimes have momentous consequences.

This chapter examines the statistical and graphical reasoning used in making two life-and-death decisions: how to stop a cholera epidemic in London during September 1854; and whether to launch the space shuttle Challenger on January 28, 1986. By creating statistical graphics that revealed the data, Dr. John Snow was able to discover the cause of the epidemic and bring it to an end. In contrast, by fooling around with displays that obscured the data, those who decided to launch the space shuttle got it wrong, terribly wrong. For both cases, the consequences resulted directly from the *quality* of methods used in displaying and assessing quantitative evidence.

The Cholera Epidemic in London, 1854

In a classic of medical detective work, *On the Mode of Communication of Cholera*,¹ John Snow described—with an eloquent and precise language of evidence, number, comparison—the severe epidemic:

The most terrible outbreak of cholera which ever occurred in this kingdom, is probably that which took place in Broad Street, Golden Square, and adjoining streets, a few weeks ago. Within two hundred and fifty yards of the spot where Cambridge Street joins Broad Street, there were upwards of five hundred fatal attacks of cholera in ten days. The mortality in this limited area probably equals any that was ever caused in this country, even by the plague; and it was much more sudden, as the greater number of cases terminated in a few hours. The mortality would undoubtedly have been much greater had it not been for the flight of the population. Persons in furnished lodgings left first, then other lodgers went away, leaving their furniture to be sent for. . . . Many houses were closed altogether owing to the death of the proprietors; and, in a great number of instances, the tradesmen who remained had sent away their families; so that in less than six days from the commencement of the outbreak, the most afflicted streets were deserted by more than three-quarters of their inhabitants.²

¹ John Snow, *On the Mode of Communication of Cholera* (London, 1855). An acute disease of the small intestine, with severe watery diarrhea, vomiting, and rapid dehydration, cholera has a fatality rate of 50 percent or more when untreated. With the rehydration therapy developed in the 1960s, mortality can be reduced to less than one percent. Epidemics still occur in poor countries, as the bacterium *Vibrio cholerae* is distributed mainly by water and food contaminated with sewage. See Dhiman Barua and William B. Greenough III, eds., *Cholera* (New York, 1992); and S. N. De, *Cholera: Its Pathology and Pathogenesis* (Edinburgh, 1961).

² Snow, *Cholera*, 38. See also *Report on the Cholera Outbreak in the Parish of St. James's, Westminster, during the Autumn of 1854*, presented to the Vestry by The Cholera Inquiry Committee (London, 1855); and H. Harold Scott, *Some Notable Epidemics* (London, 1934).

Cholera broke out in the Broad Street area of central London on the evening of August 31, 1854. John Snow, who had investigated earlier epidemics, suspected that the water from a community pump-well at Broad and Cambridge Streets was contaminated. Testing the water from the well on the evening of September 3, Snow saw no suspicious impurities, and thus he hesitated to come to a conclusion. This absence of evidence, however, was not evidence of absence:

Further inquiry . . . showed me that there was no other circumstance or agent common to the circumscribed locality in which this sudden increase of cholera occurred, and not extending beyond it, except the water of the above mentioned pump. I found, moreover, that the water varied, during the next two days, in the amount of organic impurity, visible to the naked eye, on close inspection, in the form of small white, flocculent [loosely clustered] particles. . . .³

From the General Register Office, Snow obtained a list of 83 deaths from cholera. When plotted on a map, these data showed a close link between cholera and the Broad Street pump. Persistent house-by-house, case-by-case detective work had yielded quite detailed evidence about a possible cause-effect relationship, as Snow made a kind of streetcorner correlation:

On proceeding to the spot, I found that nearly all of the deaths had taken place within a short distance of the pump. There were only ten deaths in houses situated decidedly nearer to another street pump. In five of these cases the families of the deceased persons informed me that they always sent to the pump in Broad Street, as they preferred the water to that of the pump which was nearer. In three other cases, the deceased were children who went to school near the pump in Broad Street. Two of them were known to drink the water; and the parents of the third think it probable that it did so. The other two deaths, beyond the district which this pump supplies, represent only the amount of mortality from cholera that was occurring before the irruption took place.

With regard to the deaths occurring in the locality belonging to the pump, there were sixty-one instances in which I was informed that the deceased persons used to drink the pump-water from Broad Street, either constantly or occasionally. In six instances I could get no information, owing to the death or departure of every one connected with the deceased individuals; and in six cases I was informed that the deceased persons did not drink the pump-water before their illness.⁴

Thus the theory implicating the particular pump was confirmed by the observed covariation: in this area of London, there were few occurrences of cholera exceeding the normal low level, except among those people who drank water from the Broad Street pump. It was now time to act; after all, the reason we seek causal explanations is in order to *intervene*, to govern the cause so as to govern the effect: “Policy-thinking is and must be causality-thinking.”⁵ Snow described his findings to the authorities responsible for the community water supply, the Board of Guardians of St. James’s Parish, on the evening of September 7, 1854. The Board ordered that the pump-handle on the Broad Street well be removed immediately. The epidemic soon ended.

³ Snow, *Cholera*, 39. A few weeks after the epidemic, Snow reported his results in a first-person narrative, more like a laboratory notebook or a personal journal than a modern research paper with its pristine, reconstructed science. Postmodern research claims to have added some complexities to the story of John Snow; see Howard Brody, *et al.*, “Map-Making and Myth-Making in Broad Street: The London Cholera Epidemic, 1854,” *The Lancet* 356 (July 1, 2000), 64-68.

⁴ Snow, *Cholera*, 39-40.

⁵ Robert A. Dahl, “Cause and Effect in the Study of Politics,” in Daniel Lerner, ed., *Cause and Effect* (New York, 1965), 88. Wold writes “A frequent situation is that description serves to maintain some *modus vivendi* (the control of an established production process, the tolerance of a limited number of epidemic cases), whereas explanation serves the purpose of *reform* (raising the agricultural yield, reducing the mortality rates, improving a production process). In other words, description is employed as an aid in the human *adjustment* to conditions, while explanation is a vehicle for ascendancy over the environment.” Herman Wold, “Causal Inference from Observational Data,” *Journal of the Royal Statistical Society*, A, 119 (1956), 29.

Moreover, the result of this intervention (a before/after experiment of sorts) was consistent with the idea that cholera was transmitted by impure water. Snow's explanation replaced previously held beliefs that cholera spread through the air or by some other means. In those times many years before the discovery of bacteria, one fantastic theory speculated that cholera vaporously rose out of the burying grounds of plague victims from two centuries earlier.⁶ In 1886 the discovery of the bacterium *Vibrio cholerae* confirmed Snow's theory. He is still celebrated for establishing the mode of cholera transmission *and* consequently the method of prevention: keep drinking water, food, and hands clear of infected sewage. Today at the old site of the Broad Street pump there stands a public house (a bar) named after John Snow, where one can presumably drink more safely than 140 years ago.



⁶ H. Harold Scott, *Some Notable Epidemics* (London, 1934), 3-4.

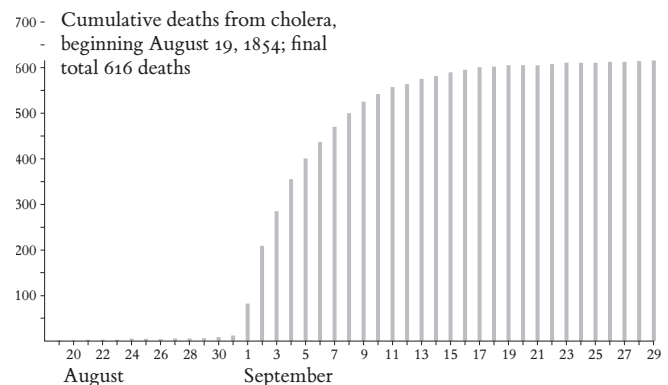
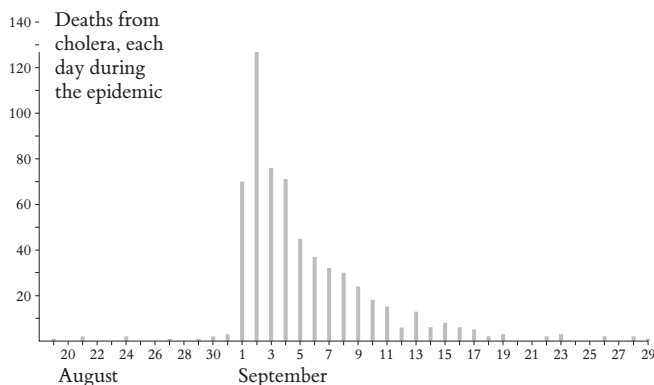
WHY was the centuries-old mystery of cholera finally solved? Most importantly, Snow had a *good idea*—a causal theory about how the disease spread—that guided the gathering and assessment of evidence. This theory developed from medical analysis and empirical observation; by mapping earlier epidemics, Snow detected a link between different water supplies and varying rates of cholera (to the consternation of private water companies who anonymously denounced Snow's work). By the 1854 epidemic, then, the intellectual framework was in place, and the problem of how cholera spread was ripe for solution.⁷

Along with a good idea and a timely problem, there was a *good method*. Snow's scientific detective work exhibits a shrewd intelligence about evidence, a clear logic of data display and analysis:

1. *Placing the data in an appropriate context for assessing cause and effect.*

The original data listed the victims' names and described their circumstances, all in order by date of death. Such a stack of death certificates naturally lends itself to time-series displays, chronologies of the epidemic as shown below. *But descriptive narration is not causal explanation*; the passage of time is a poor explanatory variable, practically useless in discovering a strategy of how to intervene and stop the epidemic.

⁷ Scientists are not “admired for failing in the attempt to solve problems that lie beyond [their] competence. . . . If politics is the art of the possible, research is surely the art of the soluble. Both are immensely practical-minded affairs. . . . The art of research [is] the art of making difficult problems soluble by devising means of getting at them. Certainly good scientists study the most important problems they think they can solve. It is, after all, their professional business to solve problems, not merely to grapple with them. The spectacle of a scientist locked in combat with the forces of ignorance is not an inspiring one if, in the outcome, the scientist is routed. That is why so many of the most important biological problems have not yet appeared on the agenda of practical research.” Peter Medawar, *Pluto's Republic* (New York, 1984), 253-254; 2-3.



Instead of plotting a time-series, which would simply report each day's bad news, Snow constructed a graphical display that provided direct and powerful testimony about a possible cause-effect relationship. Recasting the original data from their one-dimensional temporal ordering into a two-dimensional spatial comparison, Snow marked deaths from cholera (▣) on this map, along with locations of the area's 13 community water pump-wells (⊙). The notorious well is located amid an intense cluster of deaths, near the D in BROAD STREET. This map reveals a strong association between cholera and proximity to the Broad Street pump, in a context of simultaneous comparison with other local water sources and the surrounding neighborhoods without cholera.



2. *Making quantitative comparisons.* The deep, fundamental question in statistical analysis is *Compared with what?* Therefore, investigating the experiences of the victims of cholera is only part of the search for credible evidence; to understand fully the cause of the epidemic also requires an analysis of those who *escaped* the disease. With great clarity, the map presented several intriguing clues for comparisons between the living and the dead, clues strikingly visible at a brewery and a workhouse (tinted yellow here). Snow wrote in his report:

There is a brewery in Broad Street, near to the pump, and on perceiving that no brewer's men were registered as having died of cholera, I called on Mr. Huggins, the proprietor. He informed me that there were above seventy workmen employed in the brewery, and that none of them had suffered from cholera—at least in severe form—only two having been indisposed, and that not seriously, at the time the disease prevailed. The men are allowed a certain quantity of malt liquor, and Mr. Huggins believes they do not drink water at all; and he is quite certain that the workmen never obtained water from the pump in the street. There is a deep well in the brewery, in addition to the New River water. (p. 42)

Saved by the beer! And at a nearby workhouse, the circumstances of non-victims of the epidemic provided important and credible evidence about the cause of the disease, as well as a quantitative calculation of an expected rate of cholera compared with the actual observed rate:

The Workhouse in Poland Street is more than three-fourths surrounded by houses in which deaths from cholera occurred, yet out of five-hundred-thirty-five inmates only five died of cholera, the other deaths which took place being those of persons admitted after they were attacked. The workhouse has a pump-well on the premises, in addition to the supply from the Grand Junction Water Works, and the inmates never sent to Broad Street for water. If the mortality in the workhouse had been equal to that in the streets immediately surrounding it on three sides, upwards of one hundred persons would have died. (p. 42)

Such clear, lucid reasoning may seem commonsensical, obvious, insufficiently technical. Yet we will soon see a tragic instance, the decision to launch the space shuttle, when this straightforward logic of statistical (and visual) comparison was abandoned by many engineers, managers, and government officials.



3. *Considering alternative explanations and contrary cases.* Sometimes it can be difficult for researchers—who both report *and* advocate their findings—to face up to threats to their conclusions, such as alternative explanations and contrary cases. Nonetheless, the credibility of a report is enhanced by a careful assessment of *all* relevant evidence, not just the evidence overtly consistent with explanations advanced by the report. The point is to get it right, not to win the case, not to sweep under the rug all the assorted puzzles and inconsistencies that frequently occur in collections of data.⁸

Both Snow’s map and the time–sequence of deaths show several apparently contradictory instances, a number of deaths from cholera with no obvious link to the Broad Street pump. And yet . . .

In some of the instances, where the deaths are scattered a little further from the rest on the map, the malady was probably contracted at a nearer point to the pump. A cabinet-maker who resided on Noel Street [some distance from Broad Street] worked in Broad Street. . . . A little girl, who died in Ham Yard, and another who died in Angel Court, Great Windmill Street, went to the school in Dufour’s Place, Broad Street, and were in the habit of drinking the pump-water. . . .⁹

In a particularly unfortunate episode, one London resident made a special effort to obtain Broad Street well water, a delicacy of taste with a side effect that unwittingly cost two lives. Snow’s report is one of careful description and precise logic:

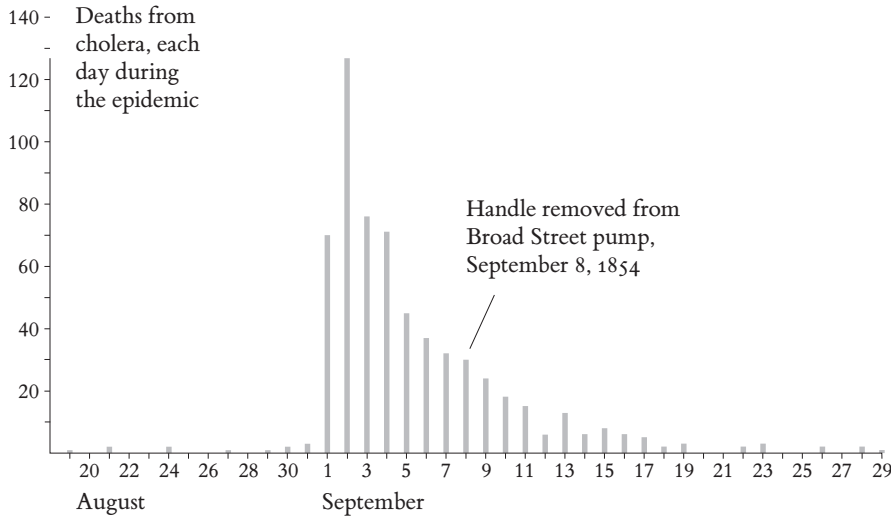
Dr. Fraser also first called my attention to the following circumstances, which are perhaps the most conclusive of all in proving the connexion between the Broad Street pump and the outbreak of cholera. In the ‘Weekly Return of Births and Deaths’ of September 9th, the following death is recorded: ‘At West End, on 2nd September, the widow of a percussion-cap maker, aged 59 years, diarrhea two hours, *cholera epidemica* sixteen hours.’ I was informed by this lady’s son that she had not been in the neighbourhood of Broad Street for many months. A cart went from Broad Street to West End every day, and it was the custom to take out a large bottle of the water from the pump in Broad Street, as she preferred it. The water was taken on Thursday, 31st August, and she drank of it in the evening, and also on Friday. She was seized with cholera on the evening of the latter day, and died on Saturday. . . . A niece, who was on a visit to this lady, also drank of the water; she returned to her residence, in a high and healthy part of Islington, was attacked with cholera, and died also. There was no cholera at the time, either at West End or in the neighbourhood where the niece died.¹⁰

Although at first glance these deaths appear unrelated to the Broad Street pump, they are, upon examination, strong evidence pointing to that well. There is here a clarity and undeniability to the link between cholera and the Broad Street pump; only such a link can account for what would otherwise be a mystery, this seemingly random and unusual occurrence of cholera. And the saintly Snow, unlike some researchers, gives full credit to the person, Dr. Fraser, who actually found this crucial case.

⁸ The distinction between science and advocacy is poignantly posed when statisticians serve as consultants and witnesses for lawyers. See Paul Meier, “Damned Liars and Expert Witnesses,” and Franklin M. Fisher, “Statisticians, Econometricians, and Adversary Proceedings,” *Journal of the American Statistical Association*, 81 (1986), 269–276 and 277–286.

⁹ Snow, *Cholera*, 47.

¹⁰ Snow, *Cholera*, 44–45.



Data source: plotted from the table in Snow, *Cholera*, 49.

Ironically, the most famous aspect of Snow's work is also the most uncertain part of his evidence: it is not at all clear that the removal of the handle of the Broad Street pump had much to do with ending the epidemic. As shown by this time-series above, the epidemic was already in rapid decline by the time the handle was removed. Yet, in many retellings of the story of the epidemic, the pump-handle removal is *the* decisive event, the unmistakable symbol of Snow's contribution. Here is the dramatic account of Benjamin Ward Richardson:

On the evening of Thursday, September 7th, the vestrymen of St. James's were sitting in solemn consultation on the causes of the [cholera epidemic]. They might well be solemn, for such a panic possibly never existed in London since the days of the great plague. People fled from their homes as from instant death, leaving behind them, in their haste, all the mere matter which before they valued most. While, then, the vestrymen were in solemn deliberation, they were called to consider a new suggestion. A stranger had asked, in modest speech, for a brief hearing. Dr. Snow, the stranger in question, was admitted and in few words explained his view of the 'head and front of the offending.' He had fixed his attention on the Broad Street pump as the source and centre of the calamity. He advised removal of the pump-handle as the grand prescription. The vestry was incredulous, but had the good sense to carry out the advice. The pump-handle was removed, and the plague was stayed.¹¹

Note the final sentence, a declaration of cause and effect.¹² Modern epidemiologists, however, are somewhat skeptical about the evidence that links the removal of the pump-handle directly to the epidemic's end. Nonetheless, the decisive point is that ultimately John Snow got it exactly right:

John Snow, in the seminal act of modern public health epidemiology, performed an intervention that was non-randomized, that was appraised with historical controls, and that had major ambiguities in the equivocal time relationship between his removal of the handle of the Broad Street pump and the end of the associated epidemic of cholera—but he correctly demonstrated that the disease was transmitted through water, not air.¹³

¹¹ Benjamin W. Richardson, "The Life of John Snow, M.D.," foreword to John Snow, *On Chloroform and Other Anaesthetics: Their Action and Administration* (London, 1858), xx-xxi.

¹² Another example of the causal claim: "On September 8, at Snow's urgent request, the handle of the Broad Street pump was removed and the incidence of new cases ceased almost at once," E. W. Gilbert, "Pioneer Maps of Health and Disease in England," *The Geographical Journal*, 124 (1958), 174. Gilbert's assertion was repeated in Edward R. Tufte, *The Visual Display of Quantitative Information* (Cheshire, Connecticut, 1983), 24.

¹³ Alvan R. Feinstein, *Clinical Epidemiology: The Architecture of Clinical Research* (Philadelphia, 1985), 409-410. And A. Bradford Hill ["Snow—An Appreciation," *Proceedings of the Royal Society of Medicine*, 48 (1955), 1010] writes: "Though conceivably there might have been a second peak in the curve, and though almost certainly some more deaths would have occurred if the pump handle had remained in situ, it is clear that the end of the epidemic was not dramatically determined by its removal."

At a minimum, removing the pump-handle prevented a recurrence of cholera. Snow recognized several difficulties in evaluating the effect of his intervention; since most people living in central London had fled, the disease ran out of possible victims—which happened simultaneously with shutting down the infected water supply.¹⁴ The case against the Broad Street pump, however, was based on a diversity of additional evidence: the cholera map, studies of unusual instances, comparisons of the living and dead with their consumption of well water, and an idea about a mechanism of contamination (a nearby underground sewer had probably leaked into the infected well). Also, the finding that cholera was carried by water—a life-saving scientific discovery that showed how to intervene and prevent the spread of cholera—derived not only from study of the Broad Street epidemic but also from Snow’s mappings of several other cholera outbreaks in relation to the purity of community water supplies.

4. *Assessment of possible errors in the numbers reported in graphics.* Snow’s analysis attends to the sources and consequences of errors in gathering the data. In particular, the credibility of the cholera map grows out of supplemental details in the text—as image, word, and number combine to present the evidence and make the argument. Detailed comments on possible errors annotate both the map and the table, reassuring readers about the care and integrity of the statistical detective work that produced the data graphics:

The deaths which occurred during this fatal outbreak of cholera are indicated in the accompanying map, as far as I could ascertain them. There are necessarily some deficiencies, for in a few of the instances of persons who died in the hospitals after their removal from the neighbourhood of Broad Street, the number of the house from which they had been removed was not registered. The address of those who died after their removal to St. James’s Workhouse was not registered; and I was only able to obtain it, in a part of the cases, on application at the Master’s Office, for many of the persons were too ill, when admitted, to give any account of themselves. In the case also of some of the workpeople and others who contracted the cholera in this neighbourhood, and died in different parts of London, the precise house from which they had removed is not stated in the return of deaths. I have heard of some persons who died in the country shortly after removing from the neighbourhood of Broad Street; and there must, no doubt, be several cases of this kind that I have not heard of. Indeed, the full extent of the calamity will probably never be known. The deficiencies I have mentioned, however, probably do not detract from the correctness of the map as a diagram of the topography of the outbreak; for, if the locality of the few additional cases could be ascertained, they would probably be distributed over the district of the outbreak in the same proportion as the large number which are known.¹⁵

The deaths in the above table [the time-series of daily deaths] are compiled from the sources mentioned above in describing the map; but some deaths which were omitted from the map on account of the number of the house not being known, are included in the table. . . .¹⁶

¹⁴ “There is no doubt that the mortality was much diminished, as I said before, by the flight of the population, which commenced soon after the outbreak; but the attacks had so far diminished before the use of the water was stopped, that it is impossible to decide whether the well still contained the cholera poison in an active state, or whether, from some cause, the water had become free from it.” Snow, *Cholera*, 51–52.

¹⁵ Snow, *Cholera*, 45–46.

¹⁶ Snow, *Cholera*, 50.

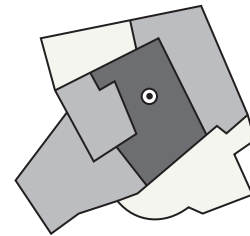
Snow drew a *dot map*, marking each individual death. This design has statistical costs and benefits: death *rates* are not shown, and such maps may become cluttered with excessive detail; on the other hand, the sometimes deceptive effects of aggregation are avoided. And of course dot maps aid in the identification and analysis of individual cases, evidence essential to Snow's argument.

The big problem is that dot maps fail to take into account the number of people living in an area and at risk to get a disease: "an area of the map may be free of cases merely because it is not populated."¹⁷ Snow's map does not fully answer the question *Compared with what?* For example, if the population as a whole in central London had been distributed just as the deaths were, then the cholera map would have merely repeated the unimportant fact that more people lived near the Broad Street pump than elsewhere. This was not the case; the entire area shown on the map—with and without cholera—was thickly populated. Still, Snow's dot map does not assess varying densities of population in the area around the pump. Ideally, the cholera data should be displayed both on a dot and a rate map, with population-based rates calculated for rather small and homogeneous geographic units. In the text of his report, however, Snow did present rates for a few different areas surrounding the pump.

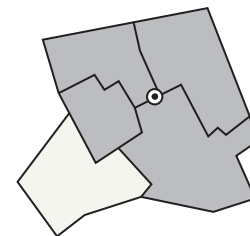
Aggregations by area can sometimes mask and even distort the true story of the data. For two of the three examples at right, constructed by Mark Monmonier from Snow's individual-level data, the intense cluster around the Broad Street pump entirely vanishes in the process of geographically aggregating the data (the greater the number of cholera deaths, the darker the area).¹⁸

In describing the discovery of how cholera is transmitted, various histories of medicine discuss the famous map and Snow's analysis. The cholera map, as Snow drew it, is difficult to reproduce on a single page; the full size of the original is awkward (a square, 40 cm or 16 inches on the side), and if reduced in size, the cholera symbols become murky and the type too small. Some facsimile editions of *On the Mode of Communication of Cholera* have given up, reprinting only Snow's text and not the crucial visual evidence of the map. Redrawings of the map for textbooks in medicine and in geography fail to reproduce key elements of Snow's original. The workhouse and brewery, those essential compared-with-what cases, are left unlabeled and unidentified, showing up only as mysterious cholera-free zones close to the infected well. Standards of quality may slip when it comes to visual displays; imprecise and undocumented work that would be unacceptable for words or tables of data too often shows up in graphics. Since it is *all* evidence—regardless of the method of presentation—the highest standards of statistical integrity and statistical thinking should apply to *every* data representation, including visual displays.

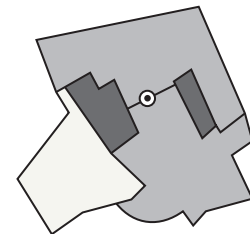
¹⁷ Brian MacMahon and Thomas F. Pugh, *Epidemiology: Principles and Methods* (Boston, 1970), 150.



In this aggregation of individual deaths into six areas, the greatest number is concentrated at the Broad Street pump.

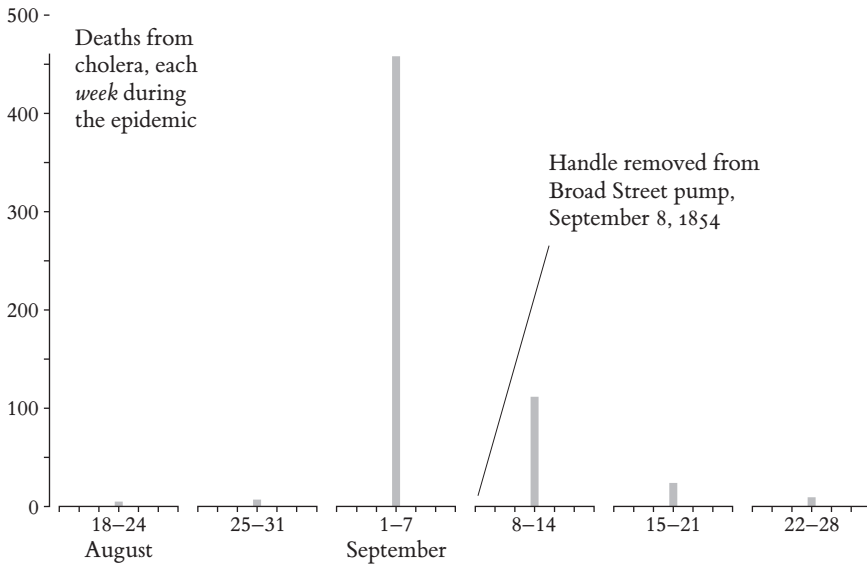
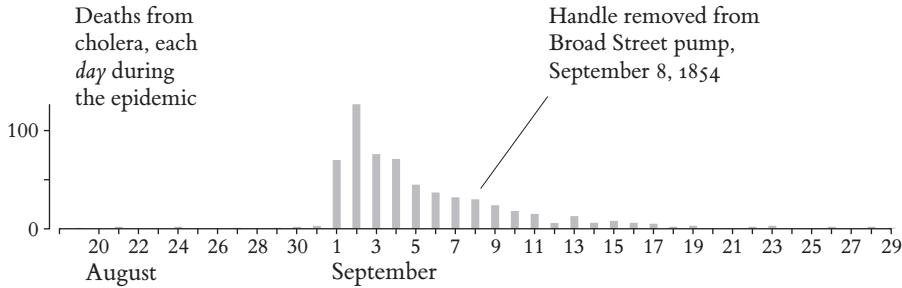


Using different geographic subdivisions, the cholera numbers are nearly the same in four of the five areas.



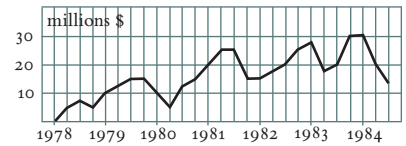
In this aggregation of the deaths, the two areas with the most deaths do not even include the infected pump!

¹⁸ Mark Monmonier, *How to Lie with Maps* (Chicago, 1991), 142-143.

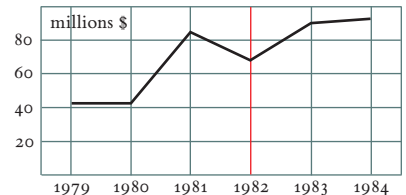


Aggregations over time may also mask relevant detail and generate misleading signals, similar to the problems of spatial aggregation in the three cholera maps. Shown at top is the familiar *daily* time-series of deaths from cholera, with its smooth decline in deaths unchanged by the removal of the pump-handle. When the daily data are added up into *weekly* intervals, however, a different picture emerges: the removal had the apparent consequence of reducing the weekly death toll from 458 to 112! But this result comes purely from the aggregation, for the daily data show no such effect.¹⁹ Conveniently, the handle was removed in early morning of September 8; hence the plausible weekly intervals of September 1-7, 8-14, and so on. Imagine if we had read the story of John Snow as reported in the first few pages here, and if our account showed the weekly instead of daily deaths—then it would all appear perfectly convincing although quite misleading.

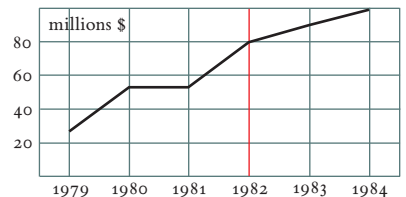
Some other weekly intervals would further aggravate the distortion. Since two or more days typically pass between consumption of the infected water and deaths from cholera, the removal date might properly be *lagged* in relation to the deaths (for example, by starting to count post-removal deaths on the 10th of September, 2 days *after* the pump-



Above, this chart shows *quarterly* revenue data in a financial graphic for a legal case. Several dips in revenue are visible.

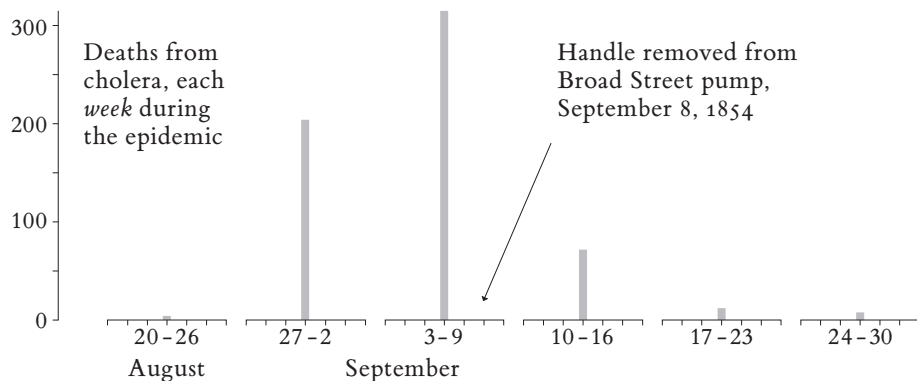


Aggregating the quarterly data into years, this chart above shows revenue by *fiscal year* (beginning July 1, ending June 30). Note the dip in 1982, the basis of a claim for damages.



Shown above are the same quarterly revenue data added up into *calendar years*. The 1982 dip has vanished.

¹⁹ Reading from the top, these clever examples reveal the effects of temporal aggregation in economic data; from Gregory Joseph, *Modern Visual Evidence* (New York, 1992), A42-A43.



handle was taken off). These lagged weekly clusters are shown above. The pseudo-effect of handle removal is now even stronger: after three weeks of increasing deaths, the weekly toll plummets when the handle is gone. A change of merely two days in weekly intervals has radically shifted the shape of the data representation. As a comparison between the two weekly charts shows, the results depend on the arbitrary choice of time periods—a sign that we are seeing method not reality.

These conjectural weekly aggregations are as condensed as news reports; missing are only the decorative clichés of “info-graphics” (the language is as ghastly as the charts). At right is how pop journalism might depict Snow’s work, complete with celebrity factoids, over-compressed data, and the isotype styling of those little coffins.

Time-series are exquisitely sensitive to choice of intervals and end points. Nonetheless, many aggregations are perfectly sensible, reducing the tedious redundancy and uninteresting complexity of large data files; for example, the *daily* data amalgamate times of death originally recorded to the hour and even minute. If in doubt, graph the detailed underlying data to assess the effects of aggregation.

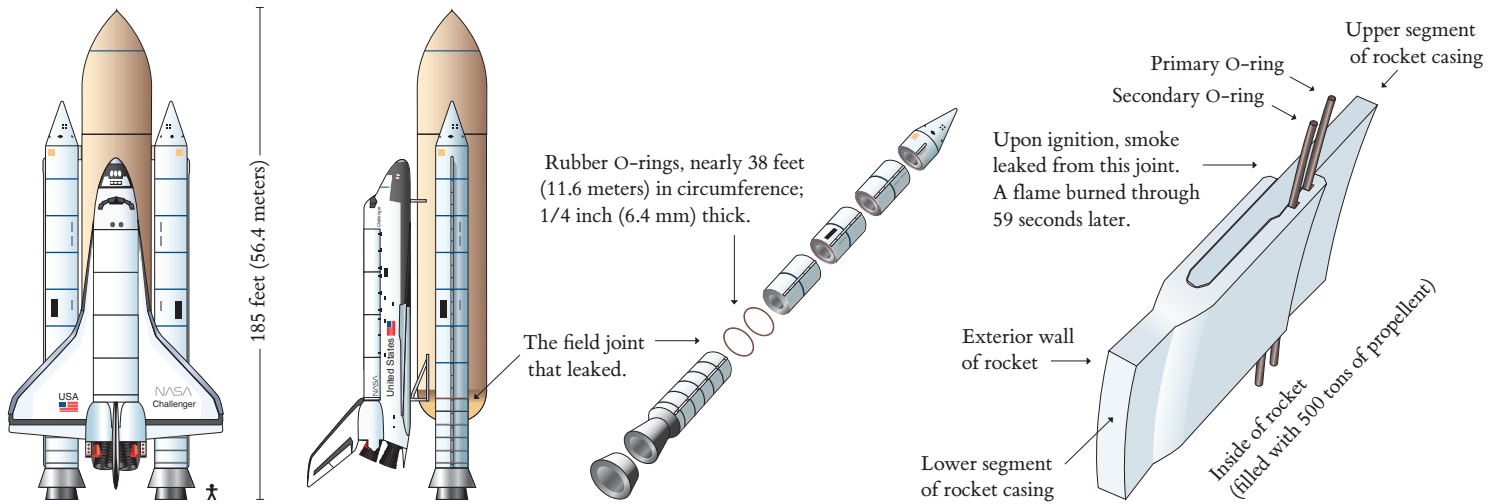
A further difficulty arises, a result of fast computing. It is easy now to sort through thousands of plausible varieties of graphical and statistical aggregations—and then to select for publication only those findings strongly favorable to the point of view being advocated. Such searches are described as *data mining*, *multiplicity*, or *specification searching*.²⁰ Thus a prudent judge of evidence might well presume that those *graphs*, *tables*, and *calculations revealed in a presentation are the best of all possible results chosen expressly for advancing the advocate’s case*.

EVEN in the face of issues raised by a modern statistical critique, it remains wonderfully true that John Snow did, after all, show exactly how cholera was transmitted and therefore prevented. In 1955, the *Proceedings of the Royal Society of Medicine* commemorated Snow’s discovery. A renowned epidemiologist, Bradford Hill, wrote: “For close upon 100 years we have been free in this country from epidemic cholera, and it is a freedom which, basically, we owe to the logical thinking, acute observations and simple sums of Dr. John Snow.”²¹



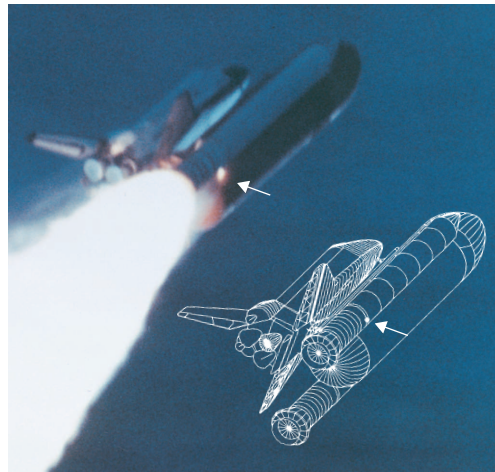
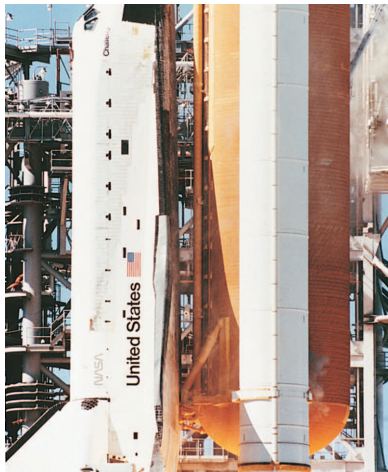
²⁰ John W. Tukey, “Some Thoughts on Clinical Trials, Especially Problems of Multiplicity,” *Science*, 198 (1977), 679–684; Edward E. Leamer, *Specification Searches: Ad Hoc Inference with Nonexperimental Data* (New York, 1978). On the other hand, “enough exploration must be done so that the results are shown to be relatively insensitive to plausible alternative specifications and data choices. Only in that way can the statistician protect himself or herself from the temptation to favor the client and from the ensuing cross-examination.” Franklin M. Fisher, “Statisticians, Econometricians, and Adversary Proceedings,” *Journal of the American Statistical Association*, 81 (1986), 279. Another reason to explore the data thoroughly is to find out what is going on! See John W. Tukey, *Exploratory Data Analysis* (Reading, Massachusetts, 1977).

²¹ A. Bradford Hill, “Snow—An Appreciation,” *Proceedings of the Royal Society of Medicine*, 48 (1955), 1012.



The shuttle consists of an *orbiter* (which carries the crew and has powerful engines in the back), a large liquid-fuel *tank* for the orbiter engines, and 2 solid-fuel *booster rockets* mounted on the sides of the central tank. Segments of the booster rockets are shipped to the launch site, where

they are assembled to make the solid-fuel rockets. Where these segments mate, each joint is sealed by two rubber O-rings as shown above. In the case of the Challenger accident, one of these joints leaked, and a torch-like flame burned through the side of the booster rocket.



Less than 1 second after ignition, a puff of smoke appeared at the aft joint of the right booster, indicating that the O-rings burned through and failed to seal. At this point, all was lost.

On the launch pad, the leak lasted only about 2 seconds and then apparently was plugged by putty and insulation as the shuttle rose, flying through rather strong cross-winds. Then 58.788 seconds after ignition, when the Challenger was 6 miles up, a flicker of flame emerged from the leaky joint. Within seconds, the flame grew and engulfed the fuel tank (containing liquid hydrogen and liquid oxygen). That tank ruptured and exploded, destroying the shuttle.



As the shuttle exploded and broke up at approximately 73 seconds after launch, the two booster rockets crisscrossed and continued flying wildly. The right booster, identifiable by its failure plume, is now to the left of its non-defective counterpart.



The flight crew of Challenger 51-L. Front row, left to right: Michael J. Smith, pilot; Francis R. (Dick) Scobee, commander; Ronald E. McNair. Back row: Ellison S. Onizuka, S. Christa McAuliffe, Gregory B. Jarvis, Judith A. Resnik.

The Decision to Launch the Space Shuttle Challenger

ON January 28, 1986, the space shuttle Challenger exploded and seven astronauts died because two rubber O-rings leaked.²² These rings had lost their resiliency because the shuttle was launched on a very cold day. Ambient temperatures were in the low 30s and the O-rings themselves were much colder, less than 20°F.

One day before the flight, the predicted temperature for the launch was 26° to 29°. Concerned that the rings would not seal at such a cold temperature, the engineers who designed the rocket opposed launching Challenger the next day. Their misgivings derived from several sources: a history of O-ring damage during previous cool-weather launches of the shuttle, the physics of resiliency (which declines exponentially with cooling), and experimental data.²³ Presented in 13 charts, this evidence was faxed to NASA, the government agency responsible for the flight. A high-level NASA official responded that he was “appalled” by the recommendation not to launch and indicated that the rocket maker, Morton Thiokol, should reconsider, even though this was Thiokol’s only no-launch recommendation in 12 years.²⁴ Other NASA officials pointed out serious weaknesses in the charts. Reassessing the situation after these skeptical responses, the Thiokol managers changed their minds and decided that they now favored launching the next day. They said the evidence presented by the engineers was inconclusive, that cool temperatures were not linked to O-ring problems.²⁵

Thus the *exact cause* of the accident was intensely debated during the evening before the launch. That is, for hours, the rocket engineers and managers considered the question: *Will the rubber O-rings fail catastrophically tomorrow because of the cold weather?* These discussions concluded at midnight with the decision to go ahead. That morning, the Challenger blew up 73 seconds after its rockets were ignited.

THE immediate cause of the accident—an O-ring failure—was quickly obvious (see the photographs at left). But what are the general causes, the lessons of the accident? And what is the meaning of Challenger? Here we encounter diverse and divergent interpretations, as the facts of the accident are reworked into moral narratives.²⁶ These allegories regularly advance claims for the special relevance of a distinct analytic approach or school of thought: if only the engineers and managers had the skills of field X, the argument implies, this terrible thing would not have happened. Or, further, the insights of X identify the deep causes of the failure. Thus, in management schools, the accident serves as a case study for reflections about groupthink, technical decision-making in the face of political pressure, and bureaucratic failures to communicate. For the authors of engineering textbooks and for the physicist Richard Feynman, the Challenger accident simply confirmed what they already

²² My sources are the five-volume *Report of the Presidential Commission on the Space Shuttle Challenger Accident* (Washington, DC, 1986) hereafter cited as *PCSSCA*; Committee on Science and Technology, House of Representatives, *Investigation of the Challenger Accident* (Washington, DC, 1986); Richard P. Feynman, “*What Do You Care What Other People Think?*” *Further Adventures of a Curious Character* (New York, 1988); Richard S. Lewis, *Challenger: The Final Voyage* (New York, 1988); Frederick Lighthall, “Launching the Space Shuttle Challenger: Disciplinary Deficiencies in the Analysis of Engineering Data,” *IEEE Transactions on Engineering Management*, 38 (February 1991), 63–74; and Diane Vaughan, *The Challenger Launch Decision: Risky Technology, Culture, and Deviance at NASA* (Chicago, 1996). The text accompanying the images at left is based on *PCSSCA*, volume 1, 6–9, 19–32, 52, 60. Illustrations of shuttle at upper left by Weilin Wu and Edward Tufte.

²³ *PCSSCA*, volume 1, 82–113.

²⁴ *PCSSCA*, volume 1, 107.

²⁵ *PCSSCA*, volume 1, 108.

²⁶ Various interpretations of the accident include *PCSSCA*, which argues several views; James L. Adams, *Flying Buttresses, Entropy, and O-Rings: The World of an Engineer* (Cambridge, Massachusetts, 1991); Michael McConnell, *Challenger: A Major Malfunction* (New York, 1987); Committee on Shuttle Criticality Review and Hazard Analysis Audit, *Post-Challenger Evaluation of Space Shuttle Risk Assessment and Management* (Washington, DC, 1988); Siddhartha R. Dalal, Edward B. Fowlkes, and Bruce Hoadley, “Risk Analysis of the Space Shuttle: Pre-Challenger Prediction of Failure,” *Journal of the American Statistical Association*, 84 (December 1989), 945–957; Claus Jensen, *No Downlink* (New York, 1996); and, cited above in note 22, the House Committee Report, the thorough account of Vaughan, Feynman’s book, and Lighthall’s insightful article.

knew: awful consequences result when heroic engineers are ignored by villainous administrators. In the field of statistics, the accident is evoked to demonstrate the importance of risk assessment, data graphs, fitting models to data, and requiring students of engineering to attend classes in statistics. For sociologists, the accident is a symptom of structural history, bureaucracy, and conformity to organizational norms. Taken in small doses, the assorted interpretations of the launch decision are plausible and rarely mutually exclusive. But when *all* these accounts are considered together, the accident appears thoroughly overdetermined. It is hard to reconcile the sense of inevitable disaster embodied in the cumulated literature of post-accident hindsight with the experiences of the first 24 shuttle launches, which were distinctly successful.

REGARDLESS of the indirect cultural causes of the accident, there was a clear proximate cause: an inability to assess the link between cool temperature and O-ring damage on earlier flights. Such a pre-launch analysis would have revealed that this flight was at considerable risk.²⁷

On the day before the launch of Challenger, the rocket engineers and managers needed a quick, smart *analysis* of evidence about the threat of cold to the O-rings, as well as an effective *presentation* of evidence in order to convince NASA officials not to launch. Engineers at Thiokol prepared 13 charts to make the case that the Challenger should *not* be launched the next day, given the forecast of very chilly weather.²⁸ Drawn up in a few hours, the charts were faxed to NASA and discussed in two long telephone conferences between Thiokol and NASA on the night before the launch. The charts were unconvincing; the arguments against the launch failed; the Challenger blew up.

These charts have weaknesses. First, the title-chart (at right, where “SRM” means Solid Rocket Motor), like the other displays, does not provide the *names* of the people who prepared the material. All too often, such documentation is absent from corporate and government reports. Public, named authorship indicates responsibility, both to the immediate audience and for the long-term record. Readers can follow up and communicate with a named source. Readers can also recall what they know about the author’s reputation and credibility. And so even a title-chart, if it lacks appropriate documentation, might well provoke some doubts about the evidence to come.

The second chart (top right) goes directly to the immediate threat to the shuttle by showing the history of eroded O-rings on launches prior to the Challenger. This varying damage, some serious but none catastrophic, was found by examining the O-rings from rocket casings retrieved for re-use. Describing the historical distribution of the *effect* endangering the Challenger, the chart does not provide data about the possible *cause*, temperature. Another impediment to understanding is that the same rocket has three different names: a NASA number (61A LH),

²⁷ The commission investigating the accident concluded: “A careful analysis of the flight history of O-ring performance would have revealed the correlation of O-ring damage and low temperature. Neither NASA nor Thiokol carried out such an analysis; consequently, they were unprepared to properly evaluate the risks of launching the 51-L [Challenger] mission in conditions more extreme than they had encountered before.” *PCSSCA*, volume I, 148. Similarly, “the decision to launch STS 51-L was based on a faulty engineering analysis of the SRM field joint seal behavior,” House Committee on Science and Technology, *Investigation of the Challenger Accident*, 10. Lighthall, “Launching the Space Shuttle,” reaches a similar conclusion.

²⁸ The 13 charts appear in *PCSSCA*, volume IV, 664-673; also in Vaughan, *Challenger Launch Decision*, 293-299.

TEMPERATURE CONCERN ON
SRM JOINTS
27 JAN 1986

HISTORY OF O-RING DAMAGE ON SRM FIELD JOINTS

SRM No.	Cross Sectional View			Top View		Clocking Location (deg)	
	Erosion Depth (in.)	Perimeter Affected (deg)	Nominal Dia. (in.)	Length Of Max Erosion (in.)	Total Heat Affected Length (in.)		
61A LH Center Field**	22A	None	None	0.280	None	None	36° -66°
61A LH Center Field**	22A	NONE	NONE	0.280	NONE	NONE	338°-18°
51C LH Forward Field**	15A	0.010	154.0	0.280	4.25	5.25	163
51C RH Center Field (prim)***	15B	0.038	130.0	0.280	12.50	58.75	354
51C RH Center Field (sec)***	15B	None	45.0	0.280	None	29.50	354
41D RH Forward Field	13B	0.028	110.0	0.280	3.00	None	275
41C LH Aft Field*	11A	None	None	0.280	None	None	--
41B LH Forward Field	10A	0.040	217.0	0.280	3.00	14.50	351
STS-2 RH Aft Field	2B	0.053	116.0	0.280	--	--	90

*Hot gas path detected in putty. Indication of heat on O-ring, but no damage.
 **Soot behind primary O-ring.
 ***Soot behind primary O-ring, heat affected secondary O-ring.

Clocking location of leak check port - 0 deg.

OTHER SRM-15 FIELD JOINTS HAD NO BLOWHOLES IN PUTTY AND NO SOOT NEAR OR BEYOND THE PRIMARY O-RING.

SRM-22 FORWARD FIELD JOINT HAD PUTTY PATH TO PRIMARY O-RING, BUT NO O-RING EROSION AND NO SOOT BLOWBY. OTHER SRM-22 FIELD JOINTS HAD NO BLOWHOLES IN PUTTY.

Thiokol's number (SRM no. 22A), and launch date (handwritten in the margin above). For O-ring damage, six types of description (erosion, soot, depth, location, extent, view) break the evidence up into stupefying fragments. An overall index summarizing the damage is needed. This chart quietly begins to define the scope of the analysis: a handful of previous flights that experienced O-ring problems.²⁹

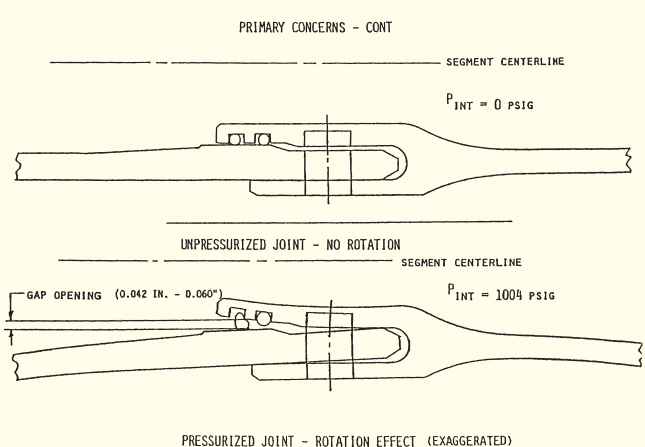
The next chart (below left) describes how erosion in the primary O-ring interacts with its back-up, the secondary O-ring. Then two drawings (below right) make an effective visual comparison to show how rotation of the field joint degrades the O-ring seal. This vital effect, however, is not linked to the potential cause; indeed, neither chart appraises the phenomena described in relation to temperature.

²⁹ This chart does not report an incident of field-joint erosion on STS 61-C, launched two weeks before the Challenger, data which appear to have been available prior to the Challenger pre-launch meeting (see PCSSCA, volume II, H-3). The damage chart is typewritten, indicating that it was prepared for an earlier presentation before being included in the final 13; handwritten charts were prepared the night before the Challenger was launched.

PRIMARY CONCERNS -

FIELD JOINT - HIGHEST CONCERN

- o EROSION PENETRATION OF PRIMARY SEAL REQUIRES RELIABLE SECONDARY SEAL FOR PRESSURE INTEGRITY
 - o IGNITION TRANSIENT - (0-600 MS)
 - o (0-170 MS) HIGH PROBABILITY OF RELIABLE SECONDARY SEAL
 - o (170-330 MS) REDUCED PROBABILITY OF RELIABLE SECONDARY SEAL
 - o (330-600 MS) HIGH PROBABILITY OF NO SECONDARY SEAL CAPABILITY
- o STEADY STATE - (600 MS - 2 MINUTES)
 - o IF EROSION PENETRATES PRIMARY O-RING SEAL - HIGH PROBABILITY OF NO SECONDARY SEAL CAPABILITY
 - o BENCH TESTING SHOWED O-RING NOT CAPABLE OF MAINTAINING CONTACT WITH METAL PARTS GAP OPENING RATE TO MEOP
 - o BENCH TESTING SHOWED CAPABILITY TO MAINTAIN O-RING CONTACT DURING INITIAL PHASE (0-170 MS) OF TRANSIENT



BLOW-BY HISTORY

SRM-15 WORST BLOW-BY

- 2 CASE JOINTS (80°), (110°) ARC
- MUCH WORSE VISUALLY THAN SRM-22

SRM 22 BLOW-BY

- 2 CASE JOINTS (30-40°)

SRM-13A, 15, 16A, 18, 23A 24A

- NOZZLE BLOW-BY

HISTORY OF O-RING TEMPERATURES
(DEGREES - F)

<u>MOTOR</u>	<u>MBT</u>	<u>AMB</u>	<u>O-RING</u>	<u>WIND</u>
DM-1	68	36	47	10 MPH
DM-2	76	45	52	10 MPH
QM-3	72.5	40	48	10 MPH
QM-4	76	48	51	10 MPH
SRM-15	52	64	53	10 MPH
SRM-22	77	78	75	10 MPH
SRM-25	55	26	29	10 MPH 25 MPH

Two charts further narrowed the evidence. Above left, “Blow-By History” mentions the two previous launches, SRM 15 and SRM 22, in which soot (blow-by) was detected in the field joints upon post-launch examination. This information, however, was already reported in the more detailed damage table that followed the title chart.³⁰ The bottom two lines refer to *nozzle* blow-by, an issue not relevant to launching the Challenger in cold weather.³¹

Although not shown in the blow-by chart, temperature is part of the analysis: SRM 15 had substantial O-ring damage and also was the coldest launch to date (at 53° on January 24, 1985, almost one year before the Challenger). This argument by analogy, made by those opposed to launching the Challenger the next morning, is reasonable, relevant, and weak. With only one case as evidence, it is usually quite difficult to make a credible statement about cause and effect.

If one case isn’t enough, why not look at two? And so the parade of anecdotes continued. By linking the blow-by chart (above left) to the temperature chart (above right), those who favored launching the Challenger spotted a weakness in the argument. While it was true that the blow-by on SRM 15 was on a cool day, the blow-by on SRM 22 was on a *warm* day at a temperature of 75° (temperature chart, second column from the right). One engineer said, “We had blow-by on the hottest motor [rocket] and on the coldest motor.”³² The superlative “-est” is an extreme characterization of these thin data, since the total number of launches under consideration here is exactly *two*.

With its focus on blow-by rather than the more common erosion, the chart of blow-by history invited the rhetorically devastating—for those opposed to the launch—comparison of SRM 15 and SRM 22. In fact, as the blow-by chart suggests, the two flights profoundly differed: the 53° launch probably barely survived with significant *erosion* of the primary and secondary O-rings on both rockets as well as blow-by; whereas the 75° launch had no erosion and only blow-by.

³⁰ On the blow-by chart, the numbers 80°, 110°, 30°, and 40° refer to the *arc* covered by blow-by on the 360° of the field (called here the “case”) joint.

³¹ Following the blow-by chart were four displays, omitted here, that showed experimental and subscale test data on the O-rings. See *PCSSCA*, volume IV, 664-673.

³² Quoted in Vaughan, *Challenger Launch Decision*, 296-297.

These charts *defined the database for the decision*: blow-by (not erosion) and temperature for two launches, SRM 15 and SRM 22. Limited measure of effect, wrong number of cases. Left out were the other 22 previous shuttle flights and their temperature variation and O-ring performance. A careful look at such evidence would have made the dangers of a cold launch clear. Displays of evidence implicitly but powerfully define the scope of the relevant, as presented data are selected from a larger pool of material. Like magicians, chartmakers reveal what they choose to reveal. That selection of data—whether partisan, hurried, haphazard, uninformed, thoughtful, wise—can make all the difference, determining the scope of the evidence and thereby setting the analytic agenda that leads to a particular decision.

For example, the temperature chart reports data for two developmental rocket motors (DM), two qualifying motors (QM), two actual launches with blow-by, and the Challenger (SRM 25) forecast.³³ These data are shown again at right. What a strange collation: the first 4 rockets were test motors that never left the ground. Missing are 92% of the temperature data, for 5 of the launches with erosion and 17 launches without erosion.

Depicting bits and pieces of data on blow-by and erosion, along with some peculiarly chosen temperatures, these charts set the stage for the unconvincing conclusions shown in two charts below. The major recommendation, “O-ring temp must be $\geq 53^\circ\text{F}$ at launch,” which was rejected, rightly implies that the Challenger could not be safely launched the next morning at 29° . Drawing a line at 53° , however, is a crudely empirical result based on a sample of size one. That anecdote was certainly not an auspicious case, because the 53° launch itself had considerable erosion. As Richard Feynman later wrote, “The O-rings of the solid rocket boosters were not designed to erode. Erosion was a clue that something was wrong. Erosion was not something from which safety could be inferred.”³⁴

³³ The table of temperature data, shown in full at left, is described as a “History of O-ring Temperatures.” It is a highly selective history, leaving out nearly all the actual flight experience of the shuttle:

MOTOR	O-RING
DM-4	47
DM-2	52
QM-3	48
QM-4	51
SRM-15	53
SRM-22	75
SRM-25	29 27

Test rockets ignited on fixed horizontal platforms in Utah.

The only 2 shuttle launches (of 24) for which temperatures were shown in the 13 Challenger charts.

Forecasted O-ring temperatures for the Challenger.

³⁴ Richard P. Feynman, “*What Do You Care What Other People Think?*” *Further Adventures of a Curious Character* (New York, 1988), 224; also in Feynman, “Appendix F: Personal Observations on the Reliability of the Shuttle,” *PCSSCA*, volume II, F2. On the many problems with the proposed 53° temperature line, see Vaughan, *Challenger Launch Decision*, 309-310.

CONCLUSIONS :

- o TEMPERATURE OF O-RING IS NOT ONLY PARAMETER CONTROLLING BLOW-BY
 SRM 15 WITH BLOW-BY HAD AN O-RING TEMP AT 53°F
 SRM 22 WITH BLOW-BY HAD AN O-RING TEMP AT 75°F
 FOUR DEVELOPMENT MOTORS WITH NO BLOW-BY WERE TESTED AT O-RING TEMP OF 47° TO 52°F
 DEVELOPMENT MOTORS HAD PUTTY PACKING WHICH RESULTED IN BETTER PERFORMANCE
- o AT ABOUT 50°F BLOW-BY COULD BE EXPERIENCED IN CASE JOINTS
- o TEMP FOR SRM 25 ON 1-28-86 LAUNCH WILL BE 29°F 9 AM
 38°F 2 PM
- o HAVE NO DATA THAT WOULD INDICATE SRM 25 IS DIFFERENT THAN SRM 15 OTHER THAN TEMP

RECOMMENDATIONS :

- o O-RING TEMP MUST BE $\geq 53^\circ\text{F}$ AT LAUNCH
 DEVELOPMENT MOTORS AT 47° TO 52°F WITH PUTTY PACKING HAD NO BLOW-BY
 SRM 15 (THE BEST SIMULATION) WORKED AT 53°F
- o PROJECT AMBIENT CONDITIONS (TEMP & WIND) TO DETERMINE LAUNCH TIME

The 13 charts failed to stop the launch. Yet, as it turned out, the chartmakers had reached the right conclusion. They had the correct theory and they were thinking causally, but they were not *displaying* causally. Unable to get a correlation between O-ring distress and temperature, those involved in the debate concluded that they didn't have enough data to quantify the effect of the cold.³⁵ The displayed data were very thin; no wonder NASA officials were so skeptical about the no-launch argument advanced by the 13 charts. For it was as if John Snow had ignored some areas with cholera and *all* the cholera-free areas and their water pumps as well. The flights without damage provide the statistical leverage necessary to understand the effects of temperature. *Numbers become evidence by being in relation to.*

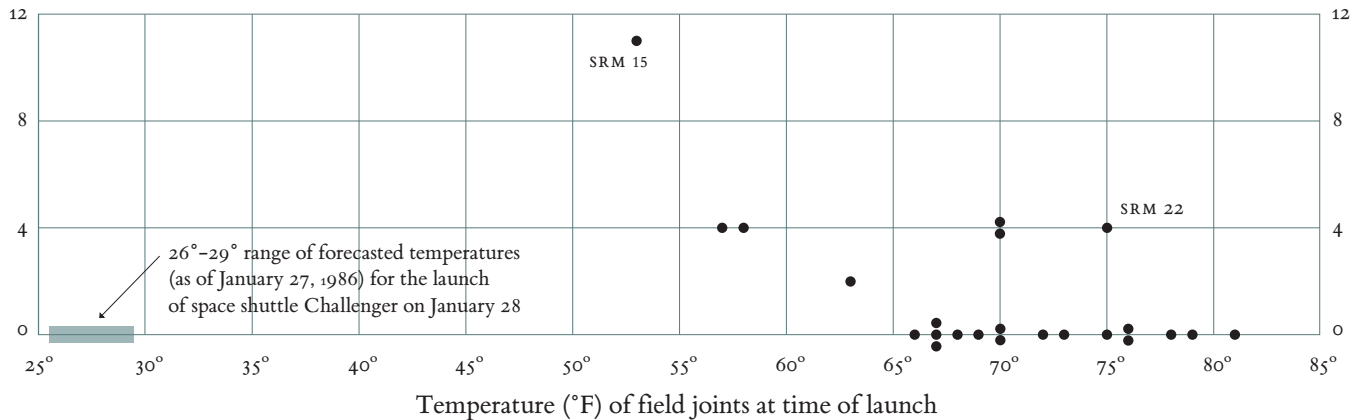
This data matrix shows the complete history of temperature and O-ring condition for all previous launches. Entries are ordered by the possible cause, temperature, from coolest to warmest launch. Data in red were exhibited at some point in the 13 pre-launch charts; and the data shown in black were not included. I have calculated an overall O-ring damage score for each launch.³⁶ The table reveals the link between O-ring distress and cool weather, with a concentration of problems on cool days compared to warm days:

³⁵ PCSSCA, volume IV, 290, 791.

³⁶ For each launch, the score on the damage index is the severity-weighted total number of incidents of O-ring erosion, heating, and blow-by. Data sources for the entire table: PCSSCA, volume II, H1-H3, and volume IV, 664; and *Post-Challenger Evaluation of Space Shuttle Risk Assessment and Management*, 135-136.

Flight	Date	Temperature °F	Erosion incidents	Blow-by incidents	Damage index	Comments
51-C	01.24.85	53°	3	2	11	Most erosion any flight; blow-by; back-up rings heated.
41-B	02.03.84	57°	1		4	Deep, extensive erosion.
61-C	01.12.86	58°	1		4	O-ring erosion on launch two weeks before Challenger.
41-C	04.06.84	63°	1		2	O-rings showed signs of heating, but no damage.
1	04.12.81	66°			0	Coollest (66°) launch without O-ring problems.
6	04.04.83	67°			0	
51-A	11.08.84	67°			0	
51-D	04.12.85	67°			0	
5	11.11.82	68°			0	
3	03.22.82	69°			0	
2	11.12.81	70°	1		4	Extent of erosion not fully known.
9	11.28.83	70°			0	
41-D	08.30.84	70°	1		4	
51-G	06.17.85	70°			0	
7	06.18.83	72°			0	
8	08.30.83	73°			0	
51-B	04.29.85	75°			0	
61-A	10.30.85	75°		2	4	No erosion. Soot found behind two primary O-rings.
51-I	08.27.85	76°			0	
61-B	11.26.85	76°			0	
41-G	10.05.84	78°			0	
51-J	10.03.85	79°			0	
	06.27.82	80°			?	O-ring condition unknown; rocket casing lost at sea.
51-F	07.29.85	81°			0	

O-ring damage
index, each launch



When assessing evidence, it is helpful to see a full data matrix, all observations for all variables, those private numbers from which the public displays are constructed. No telling what will turn up.

Above, a scatterplot shows the experience of all 24 launches prior to the Challenger. Like the table, the graph reveals the serious risks of a launch at 29°. Over the years, the O-rings had persistent problems at cooler temperatures: indeed, *every* launch below 66° resulted in damaged O-rings; on warmer days, only a few flights had erosion. In this graph, the temperature scale extends down to 29°, visually expressing the stupendous extrapolation beyond all previous experience that must be made in order to launch at 29°. The coolest flight without any O-ring damage was at 66°, some 37° warmer than predicted for the Challenger; the forecast of 29° is 5.7 standard deviations distant from the average temperature for previous launches. This launch was completely outside the engineering database accumulated in 24 previous flights.

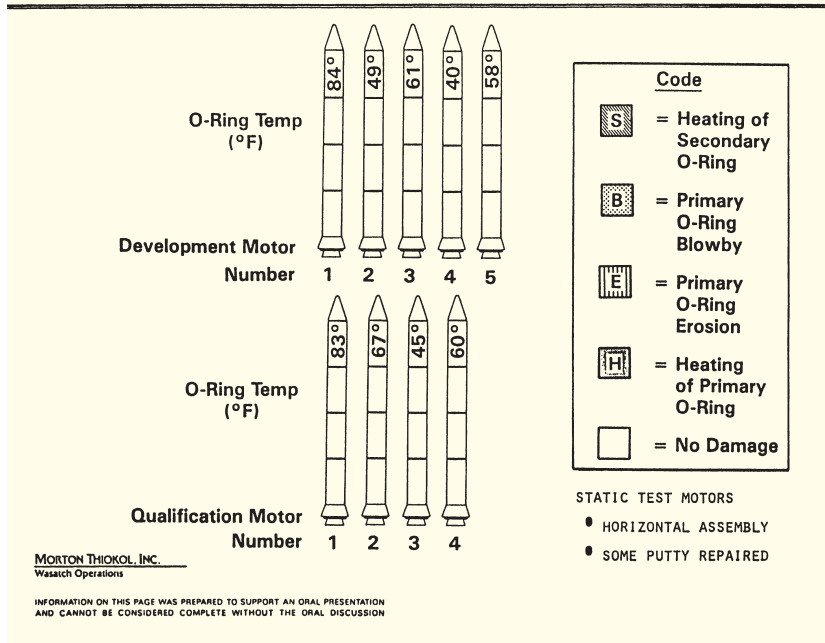
IN the 13 charts prepared for making the decision to launch, there is a scandalous discrepancy between the intellectual tasks at hand and the images created to serve those tasks. As analytical graphics, the displays failed to reveal a risk that was in fact present. As presentation graphics, the displays failed to persuade government officials that a cold-weather launch might be dangerous. In designing those displays, the chartmakers didn't quite know what they were doing, and they were doing a lot of it.³⁷ We can be thankful that most data graphics are *not* inherently misleading or uncommunicative or difficult to design correctly.

The graphics of the cholera epidemic and shuttle, and many other examples,³⁸ suggest this conclusion: *there are right ways and wrong ways to show data; there are displays that reveal the truth and displays that do not.* And, if the matter is an important one, then getting the displays of evidence right or wrong can possibly have momentous consequences.

³⁷ Lighthall concluded: "Of the 13 charts circulated by Thiokol managers and engineers to the scattered teleconferees, six contained no tabled data about either O-ring temperature, O-ring blow-by, or O-ring damage (these were primarily outlines of arguments being made by the Thiokol engineers). Of the seven remaining charts containing data either on launch temperatures or O-ring anomaly, six of them included data on either launch temperatures or O-ring anomaly but not both in relation to each other." Lighthall, "Launching the Space Shuttle Challenger," 65. See also note 27 above for the conclusions of the shuttle commission and the House Committee on Science and Technology.

³⁸ Edward R. Tufte, *The Visual Display of Quantitative Information* (Cheshire, Connecticut, 1983), 13-77.

History of O-Ring Damage in Field Joints



SOON after the Challenger accident, a presidential commission began an investigation. In evidence presented to the commission, some more charts attempted to describe the history of O-ring damage in relation to temperature. Several of these displays still didn't get it right.³⁹

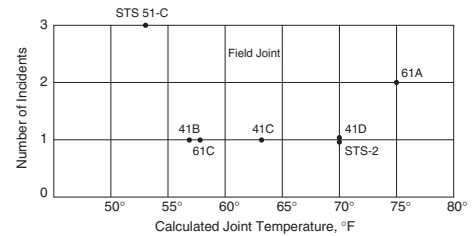
Prepared for testimony to the commission, the chart above shows nine little rockets annotated with temperature readings turned sideways. A legend shows a damage scale. Apparently measured in orderly steps, this scale starts with the most serious problem (“Heating of Secondary O-ring,” which means a primary ring burned through and leaked) and then continues in several ordered steps to “No Damage.” Regrettably, the scale’s visual representation is disordered: the cross-hatching varies erratically from dark, to light, to medium dark, to darker, to lightest—a visual pattern unrelated to the substantive order of the measured scale. A letter-code accompanies the cross-hatching. Such codes can hinder visual understanding.

At any rate, these nine rockets suffered no damage, even at quite cool temperatures. But the graph is not on point, for it is based on test data from “Development and Qualification Motors”—all fixed rockets ignited on horizontal test stands at Thiokol, never undergoing the stress of a real flight. Thus this evidence, although perhaps better than nothing (that’s all it is better than), is not directly relevant to evaluating the dangers of a cold-weather launch. Some of these same temperature numbers for test rockets are found in a pre-launch chart that we saw earlier.

Beneath the company logotype down in the lower left of this chart lurks a legalistic disclaimer (technically known as a CYA notice) that says

PCSSCA, volume v, 895.

³⁹ Most accounts of the Challenger reproduce a scatterplot that apparently demonstrates the analytical failure of the pre-launch debate. This graph depicts only launches with O-ring damage and their temperatures, omitting all damage-free launches (an absence of data points on the line of zero incidents of damage):



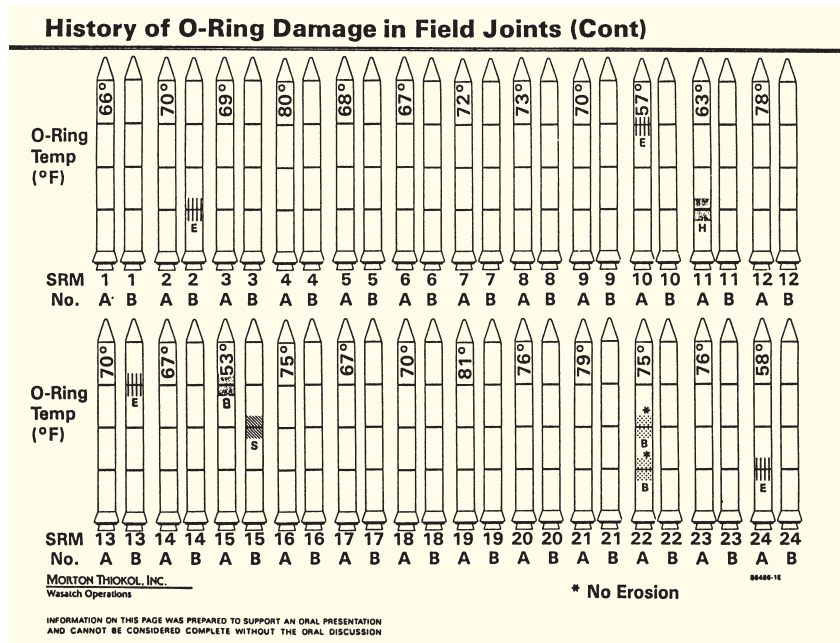
First published in the shuttle commission report (PCSSCA, volume 1, 146), the chart is a favorite of statistics teachers. It appears in textbooks on engineering, graphics, and statistics—relying on Dalal, Fowlkes, Hoadley, “Risk Analysis of the Space Shuttle: Pre-Challenger Prediction of Failure,” who describe the scatterplot as having a central role in the launch decision. (The commission report does not say when the plot was made.) The graph of the missing data-points is a vivid and poignant object lesson in how not to look at data when making an important decision. But it is too good to be true! First, the graph was *not* part of the pre-launch debate; it was *not* among the 13 charts used by Thiokol and NASA in deciding to launch. Rather, it was drawn *after* the accident by two staff members (the executive director and a lawyer) at the commission *as their simulation* of the poor reasoning in the pre-launch debate. Second, the graph implies that the pre-launch analysis examined 7 launches at 7 temperatures with 7 damage measurements. That is not true; only 2 cases of blow-by and 2 temperatures were linked up. The actual pre-launch analysis was much thinner than indicated by the commission scatterplot. Third, the damage scale is dequantified, only counting the number of incidents rather than measuring their severity. In short, whether for teaching statistics or for seeking to understand the practice of data graphics, why use an inaccurately simulated post-launch chart when we have the genuine 13 pre-launch decision charts right in hand? (On this scatterplot, see Lighthall, “Launching the Space Shuttle Challenger;” and Vaughan, *Challenger Launch Decision*, 382–384.)

this particular display should not be taken quite at face value—you had to be there:

INFORMATION ON THIS PAGE WAS PREPARED TO SUPPORT AN ORAL PRESENTATION AND CANNOT BE CONSIDERED COMPLETE WITHOUT THE ORAL DISCUSSION

Such defensive formalisms should provoke rambunctious skepticism: they suggest a corporate distrust both of the chartmaker and of any viewers of the chart.⁴⁰ In this case, the graph is documented in reports, hearing transcripts, and archives of the shuttle commission.

The second chart in the sequence is most significant. Shown below are the O-ring experiences of all 24 previous shuttle launches, with 48 little rockets representing the 24 flight-pairs:



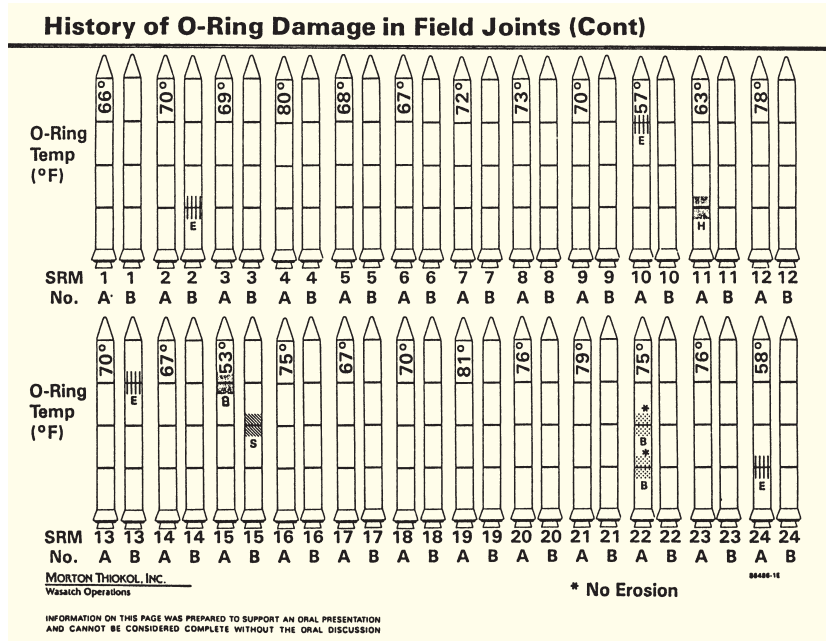
⁴⁰ This caveat, which also appeared on Thiokol's final approval of the Challenger launch (reproduced here with the epigraphs on page 26), was discussed in hearings on Challenger by the House Committee on Science and Technology: "U. Edwin Garrison, President of the Aerospace Group at Thiokol, testified that the caveat at the bottom of the paper in no way 'insinuates . . . that the document doesn't mean what it says.'" *Investigation of the Challenger Accident*, 228-229, note 80.

PCSSCA, volume v, 896.

Rockets marked with the damage code show the seven flights with O-ring problems. Launch temperature is given for each pair of rockets. Like the data matrix we saw earlier, this display contains *all* the information necessary to diagnose the relationship between temperature and damage, if we could only see it.⁴¹ The poor design makes it impossible to learn what was going on. In particular:

The Disappearing Legend At the hearings, these charts were presented by means of the dreaded overhead projector, which shows one image after another like a slide projector, making it difficult to compare and link images. When the first chart (the nine little rockets) goes away, the visual code calibrating O-ring damage also vanishes. Thus viewers need to memorize the code in order to assess the severity and type of damage sustained by each rocket in the 48-rocket chart.

⁴¹ This chart shows the rocket pair SRM 4A, SRM 4B at 80°F, as having *undamaged* O-rings. In fact, those rocket casings were lost at sea and their O-ring history is unknown.



PCSSCA, volume v, 896. This image is repeated from our page 47.

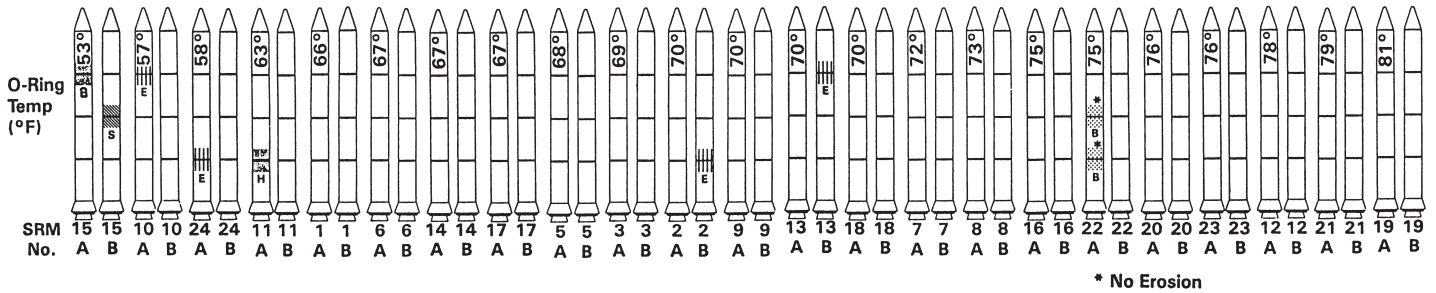
Chartjunk Good design brings *absolute attention* to data. Yet instead of focusing on a possible link between damage and temperature—the vital issue here—the strongest visual presence in this graph is the clutter generated by the outlines of the 48 little rockets. The visual elements bounce and glow, as heavy lines activate the white space, producing visual noise. Such misplaced priorities in the design of graphs and charts should make us suspicious about the competence and integrity of the analysis. Chartjunk indicates statistical stupidity, just as weak writing often reflects weak thought: “Neither can his mind be thought to be in tune, whose words do jarre,” wrote Ben Jonson in the early 1600s, “nor his reason in frame, whose sentence is preposterous.”⁴²

Lack of Clarity in Depicting Cause and Effect Turning the temperature numbers sideways obscures the causal variable. Sloppy typography also impedes inspection of these data, as numbers brush up against line-art. Likewise garbled is the measure of effect: O-ring anomalies are depicted by little marks—scattered and opaquely encoded—rather than being totaled up into a summary score of damage for each flight. Once again Jonson’s Principle: these problems are more than just poor design, for a lack of visual clarity in arranging evidence is a sign of a lack of intellectual clarity in reasoning about evidence.

Wrong Order The fatal flaw is the *ordering* of the data. Shown as a time-series, the rockets are sequenced by date of launching—from the first pair at upper left SRM No. 1 A B to the last pair at lower right 24 A B (the launch immediately prior to Challenger). The sequential order conceals the possible link between temperature and O-ring damage, thereby throwing statistical thinking into disarray. The time-series

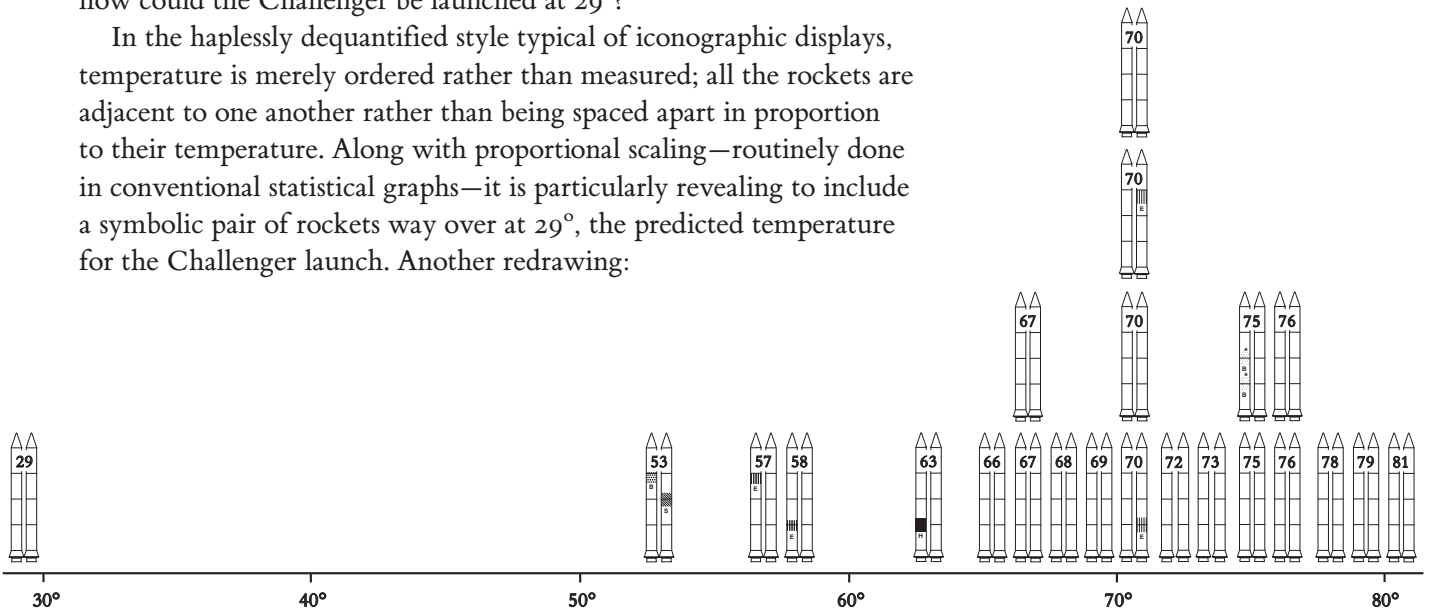
⁴² Ben Jonson, *Timber: or, Discoveries* (London, 1641), first printed in the Folio of 1640, *The Workes . . .*, p. 122 of the section beginning with *Horace his Art of Poetry*. On chartjunk, see Edward R. Tufte, *The Visual Display of Quantitative Information* (Cheshire, Connecticut, 1983), 106–121.

chart at left bears on the issue: Is there a time trend in O-ring damage?
 This is a perfectly reasonable question, but not the one on which the survival of Challenger depended. That issue was: Is there a temperature trend in O-ring damage?



Information displays should serve the analytic purpose at hand; if the substantive matter is a possible cause-effect relationship, then graphs should organize data so as to illuminate such a link. Not a complicated idea, but a profound one. Thus the little rockets must be *placed in order by temperature, the possible cause*. Above, the rockets are so ordered by temperature. This clearly shows the serious risks of a cold launch, for most O-ring damage occurs at cooler temperatures. Given this evidence, how could the Challenger be launched at 29°?

In the haplessly dequantified style typical of iconographic displays, temperature is merely ordered rather than measured; all the rockets are adjacent to one another rather than being spaced apart in proportion to their temperature. Along with proportional scaling—routinely done in conventional statistical graphs—it is particularly revealing to include a symbolic pair of rockets way over at 29°, the predicted temperature for the Challenger launch. Another redrawing:



Even after repairs, the pictorial approach with cute little rockets remains ludicrous and corrupt. The excessively original artwork just plays around with the information. It is best to forget about designs involving such icons and symbols—in this case and, for that matter, in nearly all other cases. These data require only a simple scatterplot or an ordered table to reveal the deadly relationship.



Photograph by Marilyn K. Yee, NYT Pictures, *The New York Times*.

AT a meeting of the commission investigating the shuttle accident, the physicist Richard Feynman conducted a celebrated demonstration that clarified the link between cold temperature and loss of resiliency in the rubber O-rings. Although this link was obvious for weeks to engineers and those investigating the accident, various officials had camouflaged the issue by testifying to the commission in an obscurantist language of evasive technical jargon.⁴³ Preparing for the moment during the public hearing when a piece of an O-ring (from a model of the field joint) would be passed around, Feynman had earlier that morning purchased a small clamp at a hardware store in Washington. A colorful theater of physics resulted. Feynman later described his famous experiment:

The model comes around to General Kutyna, and then to me. The clamp and pliers come out of my pocket, I take the model apart, I've got the O-ring pieces in my hand, but I still haven't got any ice water! I turn around again and signal the guy I've been bothering about it, and he signals back, "Don't worry, you'll get it!" . . .

So finally, when I get my ice water, I don't drink it! I squeeze the rubber in the C-clamp, and put them in the glass of ice water. . . .

I press the button for my microphone, and I say, "I took this rubber from the model and put it in a clamp in ice water for a while."

I take the clamp out, hold it in the air, and loosen it as I talk: "I discovered that when you undo the clamp, the rubber doesn't spring back. In other words, for more than a few seconds, there is no resilience in this particular material when it is at a temperature of 32 degrees. I believe that has some significance for our problem."⁴⁴

⁴³ One official "gave a vivid flavor of the engineering jargon—the tang end up and the clevis end down, the grit blast, the splashdown loads and cavity collapse loads, the Randolph type two zinc chromate asbestos-filled putty laid up in strips—all forbidding to the listening reporters if not to the commissioners themselves." James Gleick, *Genius: The Life and Science of Richard Feynman* (New York, 1992), 422.

⁴⁴ Richard P. Feynman, "What Do You Care What Other People Think?" *Further Adventures of a Curious Character* (New York, 1988), 151–153. Feynman's words were edited somewhat in this posthumously published book; for the actual hearings, see *PCSSCA*, volume IV, 679, transcript.



To create a more effective exhibit, the clamped O-ring might well have been placed in a transparent glass of ice water rather than in the opaque cup provided to Feynman. Such a display would then make a visual reference to the extraordinary pre-flight photographs of an ice-covered launch pad, thereby tightening up the link between the ice-water experiment and the Challenger.⁴⁵

With a strong visual presence and understated conclusion (“I believe that has some significance for our problem”), this science experiment, improvised by a Nobel laureate, became a media sensation, appearing on many news broadcasts and even on the front page of *The New York Times*. Alert to these possibilities, Feynman had deliberately provided a vivid “news hook” for an apparently inscrutable technical issue in rocket engineering:

During the lunch break, reporters came up to me and asked questions like, “Were you talking about the O-ring or the putty?” and “Would you explain to us what an O-ring is, exactly?” So I was rather depressed that I wasn’t able to make my point. But that night, all the news shows caught on to the significance of the experiment, and the next day, the newspaper articles explained everything perfectly.⁴⁶

Never have so many viewed a single physics experiment. As Freeman Dyson rhapsodized: “The public saw with their own eyes how science is done, how a great scientist thinks with his hands, how nature gives a clear answer when a scientist asks her a clear question.”⁴⁷

AND yet the presentation is deeply flawed, committing the same type of error of omission that was made in the 13 pre-launch charts. Another anecdote, without variation in cause or effect, the ice-water experiment is *uncontrolled and dequantified*. It does not address the questions *Compared with what? At what rate?* Consequently the evidence of a one-glass exhibit is equivocal: Did the O-ring lose resilience because it was clamped hard, because it was cold, or because it was wet? A credible experimental

⁴⁵ Above, icicles hang from the service structure for the Challenger. At left, the photograph shows icicles near the solid-fuel booster rocket; for a sense of scale, note that the white booster rocket is 12 ft (3.7 m) in diameter. From *PCSSCA*, volume 1, 113. One observer described the launch service tower as looking like “. . . something out of Dr. Zhivago. There’s sheets of icicles hanging everywhere.” House Committee on Science and Technology, *Investigation of the Challenger Accident*, 238. Illustration of O-ring experiment by Weilin Wu and Edward Tufte.

⁴⁶ Feynman, “What Do You Care What Other People Think?”, 153.

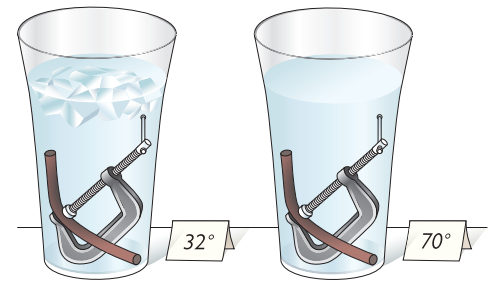
⁴⁷ Freeman Dyson, *From Eros to Gaia* (New York, 1992), 312.

design requires at least two clamps, two pieces of O-ring, and two glasses of water (one cold, one not). The idea is that the two O-ring pieces are alike in all respects save their exposure to differing temperatures. Upon releasing the clamps from the O-rings, presumably only the cold ring will show reduced resiliency. In contrast, the one-glass method is not an experiment; it is merely an experience.

For a one-glass display, neither the cause (ice water in an opaque cup) nor the effect (the clamp's imprint on the O-ring) is explicitly shown. Neither variable is quantified. In fact, neither variable varies.

A controlled experiment would not merely evoke the well-known empirical connection between temperature and resiliency, but would also reveal the overriding *intellectual* failure of the pre-launch analysis of the evidence. That failure was a lack of control, a lack of comparison.⁴⁸ The 13 pre-launch charts, like the one-glass experiment, examine only a few instances of O-ring problems and not the causes of O-ring success. A sound demonstration would exemplify the idea that in reasoning about causality, *variations in the cause* must be explicitly and measurably linked to *variations in the effect*. These principles were violated in the 13 pre-launch charts as well as in the post-launch display that arranged the 48 little rockets in temporal rather than causal order. Few lessons about the use of evidence for making decisions are more important: story-telling, weak analogies, selective reporting, warped displays, and anecdotes are not enough.⁴⁹ Reliable knowledge grows from evidence that is collected, analyzed, and displayed with some good comparisons in view. And why should we fail to be rigorous about evidence and its presentation just because the evidence is a part of a public dialogue, or is meant for the news media, or is about an important problem, or is part of making a critical decision in a hurry and under pressure?

Failure to think clearly about the analysis and the presentation of evidence opens the door for all sorts of political and other mischief to operate in making decisions. For the Challenger, there were substantial pressures to get it off the ground as quickly as possible: an unrealistic and over-optimistic flight schedule based on the premise that launches were a matter of routine (this massive, complex, and costly vehicle was named the "shuttle," as if it made hourly flights from Boston to New York); the difficulty for the rocket-maker (Morton Thiokol) to deny the demands of its major client (NASA); and a preoccupation with public relations and media events (there was a possibility of a televised conversation between the orbiting astronaut-teacher Christa McAuliffe and President Reagan during his State of the Union address that night, 10 hours after the launch). But these pressures would not have prevailed over credible evidence against the launch, for many other flights had been delayed in the past for good reasons. Had the correct scatterplot or data table been constructed, no one would have dared to risk the Challenger in such cold weather.



⁴⁸ Feynman was aware of the problematic experimental design. During hearings in the afternoon following the ice-water demonstration, he began his questioning of NASA management with this comment: "We spoke this morning about the resiliency of the seal, and if the material weren't resilient, it wouldn't work in the appropriate mode, or it would be less satisfactory, in fact, it might not work well. I did a little experiment here, and *this is not the way to do such experiments*, indicating that the stuff looked as if it was less resilient at lower temperatures, in ice." (*PCSSCA*, volume IV, 739-740, transcript, emphasis added.) Drawing of two-glass experiment by Weilin Wu and Edward Tufte.

⁴⁹ David C. Hoaglin, Richard J. Light, Bucknam McPeck, Frederick Mosteller, and Michael Stoto, *Data for Decisions: Information Strategies for Policymakers* (Cambridge, Massachusetts, 1982).

Conclusion: Thinking and Design

RICHARD Feynman concludes his report on the explosion of the space shuttle with this blunt assessment: “For a successful technology, reality must take precedence over public relations, for Nature cannot be fooled.”⁵⁰ Feynman echoes the similarly forthright words of Galileo in 1615: “It is not within the power of practitioners of demonstrative sciences to change opinion at will, choosing now this and now that one; there is a great difference between giving orders to a mathematician or a philosopher and giving them to a merchant or a lawyer; and demonstrated conclusions about natural and celestial phenomena cannot be changed with the same ease as opinions about what is or is not legitimate in a contract, in a rental, or in commerce.”⁵¹

In our cases here, the inferences made from the data faced exacting reality tests: the cholera epidemic ends or persists, the shuttle flies or fails. Those inferences and the resulting decisions and actions were based on various visual representations (maps, graphs, tables) of the evidence. The quality of these representations differed enormously, and in ways that governed the ultimate consequences.

For our case studies, and surely for the many other instances where evidence makes a difference, the conclusion is unmistakable: if displays of data are to be truthful and revealing, then the design logic of the display must reflect the intellectual logic of the analysis:

Visual representations of evidence should be governed by principles of reasoning about quantitative evidence. For information displays, design reasoning must correspond to scientific reasoning. Clear and precise seeing becomes as one with clear and precise thinking.

For example, the scientific principle, *make controlled comparisons*, also guides the construction of data displays, prescribing that the ink or pixels of graphics should be arranged so as to depict comparisons and contexts. Display architecture recapitulates quantitative thinking; design quality grows from intellectual quality. Such dual principles—both for reasoning about statistical evidence *and* for the design of statistical graphics—include (1) *documenting* the sources and characteristics of the data, (2) insistent enforcement of appropriate *comparisons*, (3) demonstrating mechanisms of *cause and effect*, (4) expressing those mechanisms *quantitatively*, (5) recognizing the inherently *multivariate* nature of analytic problems, and (6) inspecting and evaluating *alternative explanations*. When consistent with the substance and in harmony with the content, information displays should be documentary, comparative, causal and explanatory, quantified, multivariate, exploratory, skeptical.

And, as illustrated by the divergent graphical practices in our cases of the epidemic and the space shuttle, it also helps to have an endless commitment to finding, telling, and showing the truth.

⁵⁰ Richard P. Feynman, “Appendix F: Personal Observations on the Reliability of the Shuttle,” *PCSSCA* volume II, F5; also, Feynman, “*What Do You Care What Other People Think?*” *Further Adventures of a Curious Character* (New York, 1988), 237.

⁵¹ Galileo Galilei, letter to the Grand Duchess Christina of Tuscany, 1615, in *The Galileo Affair: A Documentary History*, edited and translated by Maurice A. Finocchiaro (Berkeley, 1989), 101.

A decorative graphic on the left side of the cover consists of a vertical sequence of red shapes: a diamond with a white circle, a square, a diamond with a white circle, a square, a diamond with a white circle, and a square. The shapes are arranged in a staggered, descending pattern from top-left to bottom-left.

**DATA
ANALYSIS
FOR
POLITICS
AND
POLICY**

EDWARD R. TUFTE

DATA ANALYSIS FOR POLITICS AND POLICY

EDWARD R. TUFTE

Professor Emeritus of Political Science,
Statistics, and Computer Science,
Yale University

FOUNDATIONS OF MODERN POLITICAL SCIENCE SERIES

Editor, **ROBERT A. DAHL**

Copyright © 2006 by Edward Rolf Tufte
Published by Graphics Press LLC
Post Office Box 430, Cheshire, Connecticut 06410
www.tufte.com

All rights to text and illustrations are reserved by Edward Rolf Tufte. This work may not be copied, reproduced, or translated in whole or in part without written permission of the publisher, except for brief excerpts in reviews or scholarly analysis. Use with any form of information storage and retrieval, electronic adaptation or whatever, computer software, or by similar or dissimilar methods now known or developed in the future is strictly forbidden without written permission of the publisher and copyright holder.

Originally published in 1974 by Prentice-Hall, Inc.

Contents

PREFACE

ix

CHAPTER 1

INTRODUCTION TO DATA ANALYSIS

1

Introduction,

Causal Explanation, 2

An Example:

*Do Automobile Safety Inspections
Save Lives?,* 5

*Developing Explanations for the
Observed Relationship,* 18

Costs and Unquantifiable Aspects, 29

CHAPTER 2

PREDICTIONS AND PROJECTIONS:
SOME ISSUES OF RESEARCH DESIGN

31

Introduction, 31

Problem in Prediction:

*The National Crime Test and a
Cancer Test,* 36

Election-Night Forecasting, 40

Bellwether Electoral Districts, 46

Regression Toward the Mean:

*How Prior Selection Affects the Measurement
of Future Performance,* 55

Prediction of Accident Proneness:

*Can Producers of Automobile Accidents
Be Identified in Advance as Consumers
of Traffic Violations?,* 60

Spellbinding Extrapolation, 63

CHAPTER 3

TWO-VARIABLE LINEAR REGRESSION

65

Introduction, 65

Example 1: Presidential Popularity and the Results of Congressional Elections, 73

Example 2: Lung Cancer and Smoking, 78

Example 3: Increase in the Number of Radios and Increase in the Number of Mental

Defectives, Great Britain, 1924–1937, 88

Example 4: The Relationship Between Seats and Votes in Two-Party Systems, 91

Example 5: Comparing the Slope and the Correlation Coefficient, 101

Example 6: Interpretation of Regression Coefficients When the Variables are Re-expressed as Logarithms, 108

CHAPTER 4

MULTIPLE REGRESSION

135

The Model, 135

Example 1: Midterm Congressional Elections—Presidential Popularity and Economic Conditions, 139

Example 2: Equality of Educational Opportunity and Multicollinearity, 148

Example 3: A Five-variable Regression—the Size of Democratic Parliaments, 156

APPENDIX

164

Notes On Obtaining Data And Other Information, 164

INDEX

171

Preface

This book demonstrates some statistical techniques useful in the study of politics and policy. My aim is to present fundamental material not found in statistics books, and, in particular, to show techniques of quantitative analysis in action on problems of politics and public policy. Most of the examples can be understood without a mathematical or statistical background; some sections require familiarity with basic statistical inference. Not all methodological bases are touched; still, in the chapters that follow, quite a number of important statistical concepts are illustrated.

The approach centers on fitting equations to data. More fundamental, however, is the illustration and development of good statistical thinking—a sense of judgment about what we can and can't learn about the world by looking at quantitative data.

Much of this material was first prepared for courses I have taught at Princeton University. I am indebted to several of my students and colleagues for suggesting improvements and also to Marver Bernstein who first encouraged me to teach a course in the quantitative analysis of public policy issues. I am deeply grateful to many people for their help, both direct and indirect, in the writing of this volume. In particular, John McCarthy read several drafts with great care; and Walter Gilbert, Walter Murphy, Dennis Thompson, and David Wallace commented on various sections of the manuscript. Over the years, Robert Dahl, Stanley Kelley, Jr., Frederick Mosteller, and John Tukey have given me good advice and encouragement on this project. Joseph G. Verbalis, Alice Anne Navin, Jan Juran, and Marge Cruise helped to gather and analyze much of the data. Mrs. Virginia Anderson prepared the final manuscript with care and accuracy. Barbra and Irma Kay Power provided a room of my own in London for writing the first draft. The section in Chapter 2 on bellwether electoral districts was coauthored with Richard A. Sun—and, without his energy and persistence, that difficult project would never have been completed. At Princeton University, the Computer Center, the Woodrow Wilson School, and the Department of Politics all provided superb institutional support. Finally, a fellowship at the Center for Advanced Study in the Behavioral Sciences in 1973–74 gave me time for final revisions. These individuals and institutions are not, of course, responsible for the faults of the book; they did help me very much and I am deeply indebted to them. In addition, I especially thank David Hoaglin of Harvard University for his careful reading of the first printing.

THE FAIRLY INTELLIGENT FLY

A large spider in an old house built a beautiful web in which to catch flies. Every time a fly landed on the web and was entangled in it the spider devoured him, so that when another fly came along he would think the web was a safe and quiet place in which to rest. One day a fairly intelligent fly buzzed around above the web so long without lighting that the spider appeared and said, "Come on down." But the fly was too clever for him and said, "I never light where I don't see other flies and I don't see any other flies in your house." So he flew away until he came to a place where there were a great many other flies. He was about to settle down among them when a bee buzzed up and said, "Hold it, stupid, that's flypaper. All those flies are trapped." "Don't be silly," said the fly, "they're dancing." So he settled down and became stuck to the flypaper with all the other flies.

Moral: There is no safety in numbers, or in anything else.

James Thurber,
Fables for Our Time

Introduction to Data Analysis

“Because that’s where they keep the money.”

—Willie Sutton, *when asked why he robbed banks*

Introduction

Students of political and social problems use statistical techniques to help

test theories and explanations by confronting them with empirical evidence,
summarize a large body of data into a small collection of typical values,
confirm that relationships in the data did not arise merely because of happenstance or random error,
discover some new relationship in the data, and
inform readers about what is going on in the data.

The use of statistical methods to analyze data does not make a study any more “scientific,” “rigorous,” or “objective.” The purpose of quantitative analysis is not to sanctify a set of findings. Unfortunately, some studies, in the words of one critic, “use statistics as a drunk uses a street lamp, for support rather than illumination.” Quantitative techniques will be more likely to illuminate if the data analyst is guided in methodological choices by a substantive understanding of the problem he or she is trying to learn about. Good procedures in data analysis involve techniques that help to (a) answer the substantive questions at hand, (b) squeeze all the relevant in-

formation out of the data, and (c) learn something new about the world.

Causal Explanation

All inquiry begins with a problem, a question to be answered. Why have some countries, despite great natural resources, remained economically weak? Why do some nations spend more on military equipment than others? Does smoking cause lung cancer? Do automobile safety inspections reduce the number of traffic accidents? Do economic conditions help determine what candidates the people vote for?

The thing to be explained is the *response variable* or *dependent variable*. In the questions above, the response variables are, respectively, the level of economic development, military expenditures, the frequency of lung cancer, the number of traffic accidents, and an individual's choice in an election. The causes, explanations, or predictors of the response variable are the *describing variables* or *independent variables*. Usually more than one describing variable will help explain the response variable; and an analysis with several describing variables is called, in the jargon, *multivariate analysis*. For example, two causes of lung cancer might be smoking and amount of time spent in a coal mine. Here the two describing variables are the amount of smoking and amount of time digging coal (and inhaling coal and rock dust).

Although it is sometimes difficult to speak in causal terms in studies of social problems, it is clear that if we want to explain or change anything, we will eventually have to work in terms of cause and effect. As Dahl put it, "policy-thinking is and must be causality-thinking."¹ Wold has even suggested a link between explanation and policy outcomes:

A frequent situation is that description serves to maintain some *modus vivendi* (the control of an established production process, the tolerance of a limited number of epidemic cases), whereas explanation serves the purpose of *reform* (raising the agricultural yield, reducing the mortality rates, improving a production process). In other words, description is employed as an aid in the human *adjustment* to conditions, while explanation is a vehicle for ascendancy over the environment.²

¹ Robert A. Dahl, "Cause and Effect in the Study of Politics," in Daniel Lerner, ed., *Cause and Effect* (New York: Free Press, 1965), p. 88

² Herman Wold, "Causal Inference from Observational Data," *Journal of the Royal Statistical Society, Series A*, 119 (1956), p. 29.

Sometimes, especially in studies based on data collected from observational records rather than from controlled experiments, researchers avoid causal language and use wishy-washy phrases to report their results: one variable is said to “predict” another: or a variable is “strongly related,” “associated,” or “varies regularly” with another variable. The language of association and prediction is probably most often used because the evidence seems insufficient to justify a direct causal statement. A better practice is to state the causal hypothesis and then to present the evidence along with an assessment with respect to the causal hypothesis—instead of letting the quality of the data determine the language of the explanation.

In other cases, researchers appear only interested in studying associations and have no causal mechanisms in mind. These studies seek to discover “patterns of association” and “clusters of interrelated variables.” Such discoveries can sometimes be a helpful first step toward developing explanations.

A good research design is a successful strategy for collecting and analyzing data that help to assess the validity of competing explanations of the variation in the response variable. In causal analysis, the basic purpose of research design is to observe or control covariation between the response and describing variables in a context such that these variables are not confounded with other uncontrolled or extraneous influences. Thus the key element in developing and testing explanations is *controlled comparison*. By such comparison we evaluate and decide among theories about what variables cause what effects. The importance of comparison or control groups in making inferences is illustrated by Cochran’s account of a study by Seltser and Sartwell on the effects of exposure to atomic radiation:

As pointed out by Seltser and Sartwell, the principal opportunities for investigations in human subjects are confined to the following: (a) the Japanese survivors of the atomic bombs in Hiroshima and Nagasaki, involving a single exposure, (b) groups occupationally exposed to radiation at times when the possible danger from this source was not realized—radiologists, dentists, and makers of watches with luminous dials, (c) persons who received medical radiation, as in the treatment of some forms of cancer, or infants exposed *in utero* through pelvic X-rays of the mother in the late stages of pregnancy, and (d) areas of the earth in which natural radioactivity is unusually high.

None of these sources provides more than limited material for constructing a dosage-response curve. . . .

The study by Seltser and Sartwell of the mortality of radiologists is an excellent example of the possibilities from groups occupationally or medically exposed. They chose male members of the Radiological Society of North America. For each member they obtained by a

painstaking search the status (dead or alive) as of December 31, 1958, with cause of death and any available information on other factors such as age that might influence duration of life. *Research of this type always raises the question: with what are the exposed group to be compared? Ideally, we seek a non-exposed group which is similar to the exposed group with regard to any other variable that is known or suspected to have a material effect on duration of life. . . . In an observational study the extent to which this goal can be met is of course dependent on our ability to measure such variables and to find a group that has similar distributions with respect to them.*

The authors chose two comparison groups. As the nearest to a non-exposed group they used the American Academy of Ophthalmology and Otolaryngology, whose members rarely have occasion to employ X-radiation. As an intermediate group they also included the American College of Physicians, since some of these members use X-rays, for example, in ear examinations. In such studies the inclusion of a middle group is advantageous in either adding confirmation to the results given by the two extreme groups or in casting doubt upon them. This study, however, again has the weakness that no measures of the doses of radiation experienced by the subjects are available, except as a rough guess for the group as a whole. Studies similar in structure have been done of the later development of infants *in utero*, as compared with a control group of non-exposed infants born in the same hospital at the same time.³

The importance of controlled comparison in the assessment of causal relationships is made even more bluntly in a doctor's story about the evaluation of surgical procedures:

One day when I was a junior medical student, a very important Boston surgeon visited the school and delivered a great treatise on a large number of patients who had undergone successful operations for vascular reconstruction. At the end of the lecture, a young student at the back of the room timidly asked, "Do you have any controls?" Well, the great surgeon drew himself up to his full height, hit the desk, and said, "Do you mean did I not operate on half of the patients?" The hall grew very quiet then. The voice at the back of the room very hesitantly replied, "Yes, that's what I had in mind." Then the visitor's fist really came down as he thundered, "Of course not. That would have doomed half of them to their death." God, it was quiet then, and one could scarcely hear the small voice ask, "Which half?"⁴

³William G. Cochran, "Planning and Analysis of Non-Experimental Studies," ONR Technical Report No. 19 (April 1968), Department of Statistics, Harvard University, pp. 7-9, italics added. The cited study is R. Seltser and P. E. Sartwell, "The Influence of Occupational Exposure to Radiation on the Mortality of American Radiologists and Other Medical Specialists," *American Journal of Epidemiology*, 81 (1965), 2-22.

⁴Dr. E. E. Peacock, Jr., Chairman of Surgery, University of Arizona College of Medicine; quoted in *Medical World News* (September 1, 1972), p. 45. I am indebted to my colleague Herman Somers for pointing out this citation to me.

One final point about the relationship between causal inferences and statistical analysis. Statistical techniques do not solve any of the common-sense difficulties about making causal inferences. Such techniques may help organize or arrange the data so that the numbers speak more clearly to the question of causality—but that is all statistical techniques can do. All the logical, theoretical, and empirical difficulties attendant to establishing a causal relationship persist no matter what type of statistical analysis is applied. “There is,” as Thurber moralized, “no safety in numbers, or in anything else.”

An Example: Do Automobile Safety Inspections Save Lives?

Let us now go through an example, analyzing some data to answer a particular question and, in the process, showing several basic techniques for looking at a collection of data. We will, in this example, try to find out whether compulsory automobile safety inspections (the describing variable) help reduce traffic fatalities (the response variable).

In 1967, nineteen states in the United States had some form of automobile safety inspection with the consequent correction of mechanical defects. Some states, such as New Jersey, had rather thorough yearly inspections, testing headlight alignment, other lights, brakes, steering, and tires. Other states had superficial inspections; most had none at all.

Inspections can produce significant benefits if they help to reduce the yearly toll of 55,000 deaths and 4.4 million minor and major injuries resulting from automobile crashes. The economic costs, too, are considerable: “A disproportionate number of the persons killed or permanently disabled represents an almost complete loss on a heavy investment: they are persons with twenty years of nurture behind them and presumed forty years of productive work ahead. The cost estimates are surpassingly fuzzy, but something like 2 percent of the Gross National Product seems about right, if property damage accidents are included.”⁵ Finally, one estimate is that “perhaps 20 percent of the automobile industry is required to replace or repair damaged vehicles.”⁶

But inspections also have significant costs, both of administration and enforcement as well as of delay and aggravation to the individual driver, who must often spend several hours having his car examined.

⁵Daniel P. Moynihan, “The War Against the Automobile,” *The Public Interest*, no. 3 (Spring 1966), p. 10.

⁶*Ibid.*, p. 13

Inspections cost directly about \$500 million each year—plus the hidden and nonfinancial costs to the individual driver. There are good reasons, then, for trying to find out whether inspections make any difference. If they do actually reduce the death rate significantly, inspection programs should be strengthened; if they have little effect, then the money might be better spent some other way.

We can imagine a controlled experiment—first choosing randomly a large number of cars, inspecting them and correcting their mechanical defects, and then following their history of accidents for several years. Another group of cars, remaining uninspected, would serve as a comparison or control group. Such an experiment would require a rather large sample, since fatal auto crashes are a relatively rare event, with about one car in a thousand being involved in a fatal accident in a given year. (Many cars during their lifetime, however, are in some sort of accident and probably at least one car in three winds up with blood on it.⁷)

Not only would the sample have to be large, but it would have to be randomly chosen. We couldn't rely entirely on volunteers, because those car owners who volunteered to have their cars inspected and to participate in the experiment would be likely to be quite different from the typical car owner. The more safety-conscious driver who owned a car with few mechanical defects would probably be more likely to volunteer than the owner of a dilapidated car. And so we would have to take steps to avoid a bias toward safety-conscious drivers, for they would probably be overrepresented in a volunteer group and other types of drivers underrepresented.

Unfortunately, few such social experiments of this type have ever been tried. Donald T. Campbell points out in his paper "Reforms as Experiments" that "The United States and other modern nations should be ready for an experimental approach to social reform, an approach in which we try out new programs designed to cure specific social problems, in which we learn whether or not these programs are effective, and in which we retain, imitate, modify or discard them on the basis of apparent effectiveness. . . . [M]ost ameliorative programs end up with *no* interpretable evaluation."⁸

What are some alternatives to a large-scale experiment—which would be the most inferentially sound way to study the problem—in order to evaluate the impact, if any, of automobile inspections? Two other methods provide help. First, a *time-series analysis* follows the trend of the death rate before and then after the adoption of inspections

⁷ *Ibid.*

⁸ *American Psychologist*, 24 (1969), p. 409.

in a given state. In other words, for each of the states that now have inspections, the job is to see whether fatalities decreased after the inspections were started. The states that still do not have inspections can be used as a comparison or control group to test other explanations (other than introduction of inspections) for changes in the death rate over time. Thus the control group helps us find out whether the fatality rate goes down, relative to similar states, when inspections are introduced in a given state.⁹

The second method, a *cross-section analysis*, compares at a given point in time the death rates in those states that have inspections with the death rates in those states without inspections. The important assumption here is that other factors affecting the death rate are equal for the inspected and the uninspected states. "Other things being equal" is sometimes only a faint hope, although often we can insure that at least some important things are approximately equal.

The remainder of this chapter consists of a cross-section analysis of the effects of inspections. The purpose is to show some basic concepts of data analysis by means of a substantive example. In the cross-section approach, the question becomes: "Do states that have automobile safety inspections have lower fatality rates than those states without inspections—other things being equal?" Comparing the variations in rates between inspected and uninspected states is not a perfect test—partly because both inspected and uninspected cars can cross state lines and be involved in accidents in other states. Furthermore, inspections may constitute part of a larger safety program that includes strong checks on drunken driving, better roads, and so forth. Thus, it might be more appropriate to attribute differences in death rates to an overall safety program in the state rather than just to inspections.

In summary, even if rates are low in inspected compared to uninspected states, we want to be very careful in attributing variations in rates only to the presence or absence of inspections. These and other complicating factors work against getting a clean test of the relationship between inspections and death rates. Such confounding factors enter into almost every analysis of social and political problems.

(WARNING: Typically, data analysis is messy, and little details clutter it. Not only confounding factors, but also deviant cases, minor problems in measurement, and ambiguous results lead to frustration and discouragement, so that more data are collected than analyzed. Ne-

⁹A good example of such a study is Donald T. Campbell and H. Laurence Ross, "The Connecticut Crackdown on Speeding: Time-Series Data in Quasi-Experimental Analysis," *Law and Society Review*, 3 (August 1968), 33-53, and reprinted in Edward R. Tufte, ed., *The Quantitative Analysis of Social Problems* (Reading, Mass.: Addison-Wesley, 1970) pp. 110-25.

glecting or hiding the messy details of the data reduces the researcher's chances of discovering something new. One common error is to underestimate the time necessary for the analysis. Although there is a good deal of variability, in many projects the analysis and synthesis of the data consume 80 to 90 percent of the total time spent. Often, after the initial collection and first analysis of the data, it is necessary and wise to go back and acquire additional information suggested by the first results. A good rule of thumb for deciding how long the analysis of the data actually will take is

(1) to add up all the time for *everything* you can think of—editing the data, checking for errors, calculating various statistics, thinking about the results, going back to the data to try out a new idea, and

(2) then multiply the estimate obtained in this first step by five. With these words of warning, let us get on with the present analysis).

The fifty states differ greatly in their automobile fatality rates: Connecticut, with the lowest rate, had 14.8 deaths per 100,000 residents in 1968; Wyoming, the highest, had a rate more than three times greater at 52.1 deaths per 100,000 people.¹⁰ Figure 1-1 reveals the wide variation in death rates for the states. If all states had a death rate as low as Connecticut, instead of 55,000 deaths in automobile accidents each year, only 30,000 deaths would occur—a reduction of 46 percent.

Figure 1-1, below, shows a cluster of three states with rather high rates: Wyoming, Nevada, and New Mexico all have rates near 50. Three other states—Idaho, Arizona, and Montana—also are quite high with rates exceeding 40 deaths per 100,000 people per year.

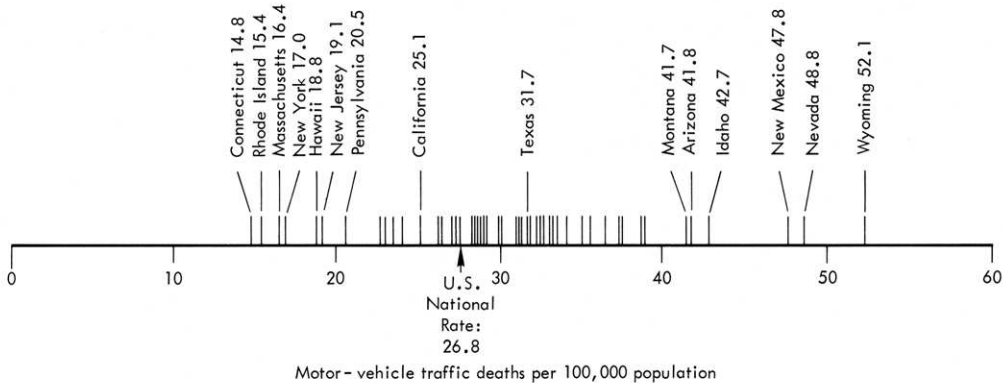


FIGURE 1-1 Death rate, motor-vehicle accidents, 1968

Six states distinguish themselves at the low end of the scale: Connecticut, Rhode Island, Massachusetts, New York, Hawaii, and New Jersey all have rates less than 20. Already, perhaps, we can see some characteristics of high-rate compared to low-rate states:

States with extremely high rates are more likely

- to be located in the western part of the United States
- to be thinly populated (i.e., low density, few people per square mile)
- not to have been one of the original 13 states of the United States

States with extremely low rates are more likely

- to be located in the eastern part of the United States
- to be thickly populated (i.e., high density, many people per square mile)
- to have been one of the original 13 states of the United States

¹⁰ Accident rates, unless otherwise noted, are taken from the appropriate annual edition of *Accidents Facts* (Chicago: National Safety Council).

States with extremely high rates are more likely

-
- not to have inspections
 - to have seven or less letters in their names

States with extremely low rates are more likely

-
- to have inspections
 - to have more than seven letters in their names

A number of factors, of varying relevance to be sure, seem to be associated with the death rate for the extremely high and extremely low states. Note that while we observe many different associations between the death rate and other characteristics of the state, it is our substantive judgment, and not merely the observed association, that tells us density and inspections might have something to do with the death rate and that the number of letters in the name of the state has nothing to do with it.

So far we have looked only at the states with either extremely high or extremely low death rates. Such a procedure, while giving some useful indications, can also be misleading: all the data should be used, not just a fraction.

In looking at Figure 1-1, one should begin to wonder just how reliable these figures are. Perhaps Wyoming is high because a bad accident involving many deaths—such as a bus accident—occurred in 1968. In a “normal” year, would Wyoming have a lower death rate? Would a different set of states fall at the low end of the scale a year before or a year after these data were collected? Do Wyoming, New Mexico, and Nevada usually have high rates—and do Rhode Island, Connecticut, and Massachusetts usually have low rates? In short, then, how do the rates vary from year to year? These questions are good ones, because if the variation in death rates across the different states changed wildly from one year to the next, we might begin to suspect that states were merely high or low because they were “lucky” or “unlucky,” because they had a few accidents resulting in many deaths in a “bad” year.

These questions are easy to answer. A number of different approaches all produce the same result: the large differences between states in their death rates have remained relatively persistent over the years. For example, the five states with the highest death rates in 1948 also had the five highest death rates in 1958 and again in 1968. Similarly the states with the five lowest death rates in 1948 were also the five lowest in 1958; in 1968, four of these five remained among the five lowest. Figure 1-2 also gives a sharp and clear answer. This scatterplot plots each state’s 1958 death rate against its 1968 rate. The picture shows:

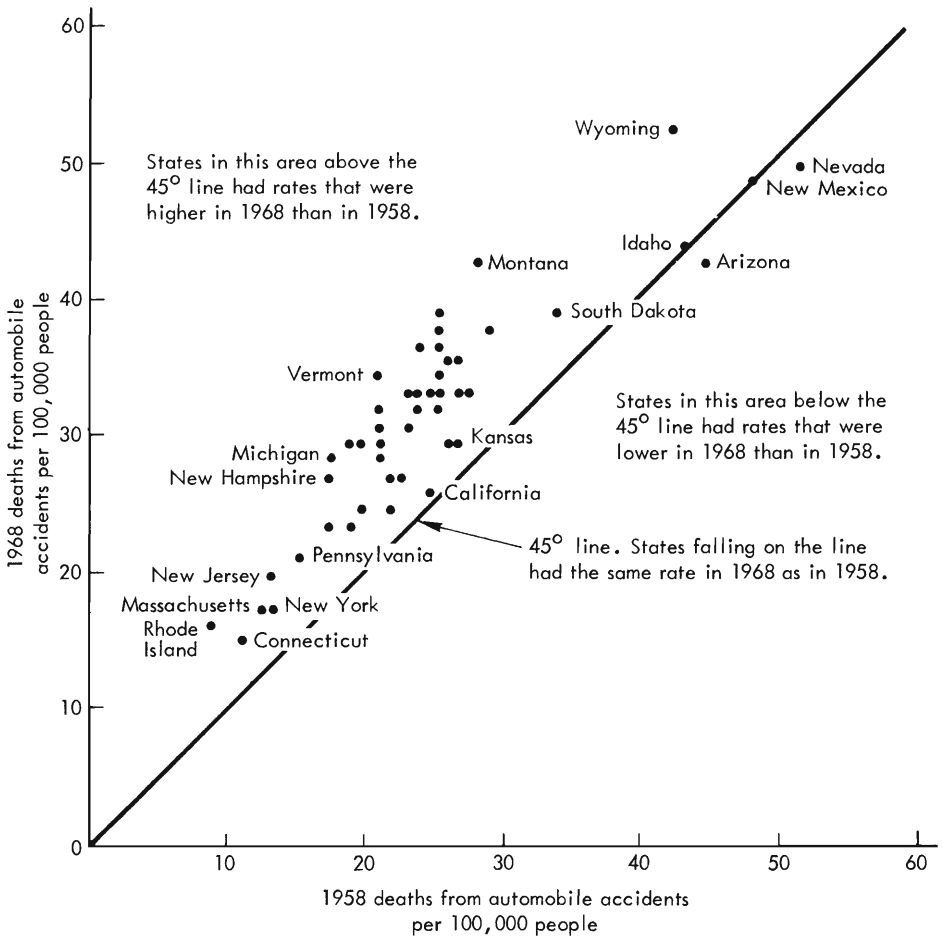


FIGURE 1-2 Death rates, 1958 and 1968

1. The states that had high rates in 1958 remained high in 1968; those with average death rates in 1958 had similar rates in 1968; and low rates in 1958 continued to be low in 1968. Such a relationship is called a positive relationship; as one variable grows bigger, so does the other variable. The scatterplot shows a fairly strong relationship in that the points increase in a relatively orderly fashion; they are not scattered all over the graph. In summary, there is a strong positive relationship between rates for 1958 and 1968.
2. Most states have somewhat higher rates in 1968 than they did ten years earlier, since most states lie *above* the 45° line (which is the area where the 1968 rate always exceeds the 1958 rate). All of the states with middle-level death rates show some increase between 1958 and 1968, since they lie above the line in the area where the 1968 rate is always greater than the 1958 rate. Finally, those states with very high death rates show a fair amount of

scatter around the 45° line, with two of them showing a lower rate in 1968 than in 1958 (since they lie *below* the 45° line).

The large differences between the various states in death rates and the relative stability of the rates over time indicate that persistent factors have great consequences for the risk one assumes when driving on the roads of the various states. The differences are not happenstance or peculiar to a particular year. There must be *something* that makes the death rate consistently three times higher in Wyoming than in Rhode Island. Since Rhode Island has safety inspections and Wyoming does not, it appears worthwhile to look into the relationship between inspections and death rates—as well as for other relationships.

Figure 1-2 shows the relative persistence of the rates for the states; the unique yearly variation does not dramatically shuffle the states relative to one another. But influences on the accident death rate peculiar or unique to a given year do contribute to some of the variation in a single year's set of accident figures for each state. In order to reduce the effect of such influences, we will average out the unique yearly variation by averaging the death rate for each state over a three-year period—with the hope of producing a fairer picture of the typical or normal behavior of the accident rate in a state. Thus, for example, the rates for Montana in 1966, 1967, and 1968 were 39.3, 45.5, and 41.7. The middle year, 1967, was unusually high and not typical of the long-run rate over the years in Montana. Yet it is an actual piece of data and not to be discounted entirely. A useful compromise, then, is the averaging technique. For Montana, the average rate over the three-year period is

$$\frac{39.3 + 45.5 + 41.7}{3} = 42.2.$$

This procedure is repeated for the remaining 49 states to compute a three-year death rate. This average rate is the response variable, the thing we are trying to explain.

Do inspections make any difference in these death rates? Figure 1-3 reveals that states with inspections tend to have lower death rates than states without inspections, although the two groups of states overlap a good deal. Most states with inspections, as the figure shows, beat the average for the uninspected states, although one inspected state, New Mexico, has an extremely high death rate compared to the rest of the inspected states. In the states without inspections, Connecticut has a very low rate (Connecticut has inspections for used cars that are sold in the state but not for new cars).

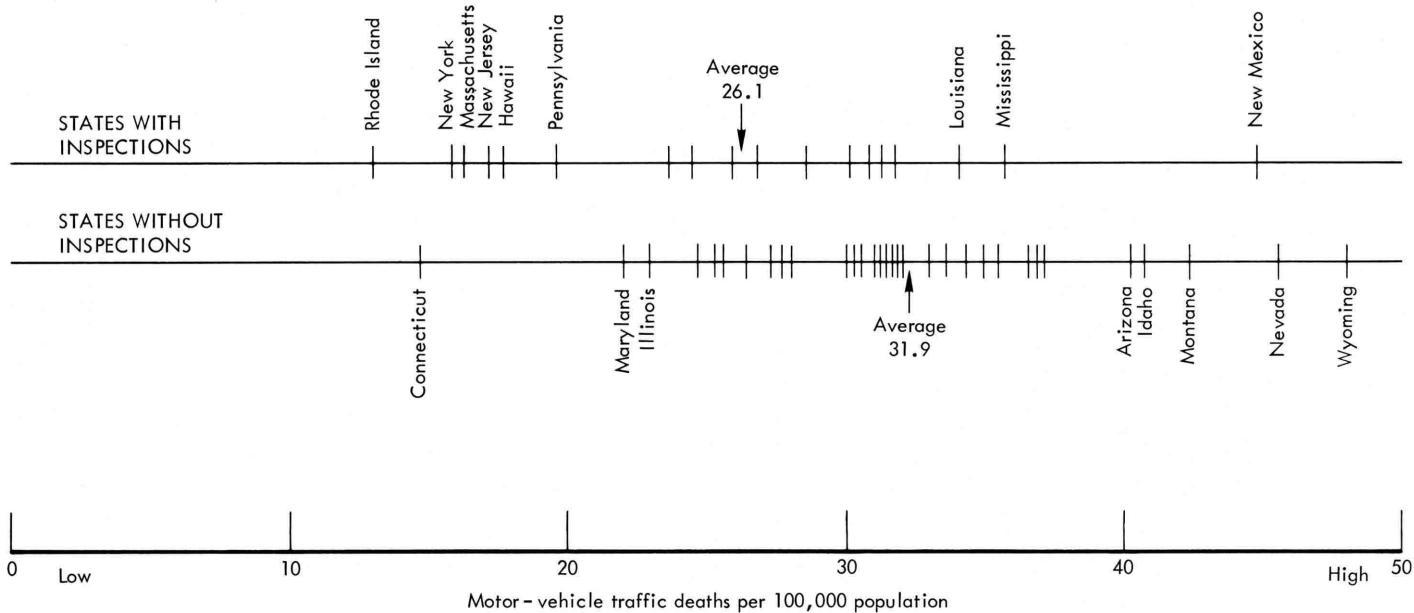


FIGURE 1-3 Inspected vs. uninspected states: averaged death rates

Figure 1-3 shows that those states with inspections typically have a death rate lower by around six deaths per 100,000 people than states without inspections. If inspections are, in fact, the cause of this observed difference, then the adoption of inspections by those states that do not have them would apparently save some 15,000 lives a year. Thus Figure 1-3, on the surface at least, indicates that inspections are very effective. But such an inference is very insecure. The most important source of doubt is that inspected and uninspected states may differ not only with respect to inspections, but also with respect to other factors that affect the death rate in automobile accidents. Thus the benefits of these other factors are wrongly attributed to inspections. (There is also a possibility that Figure 1-3 understates the benefits of inspections—for perhaps it was states with especially high death rates that adopted inspections several years ago.)

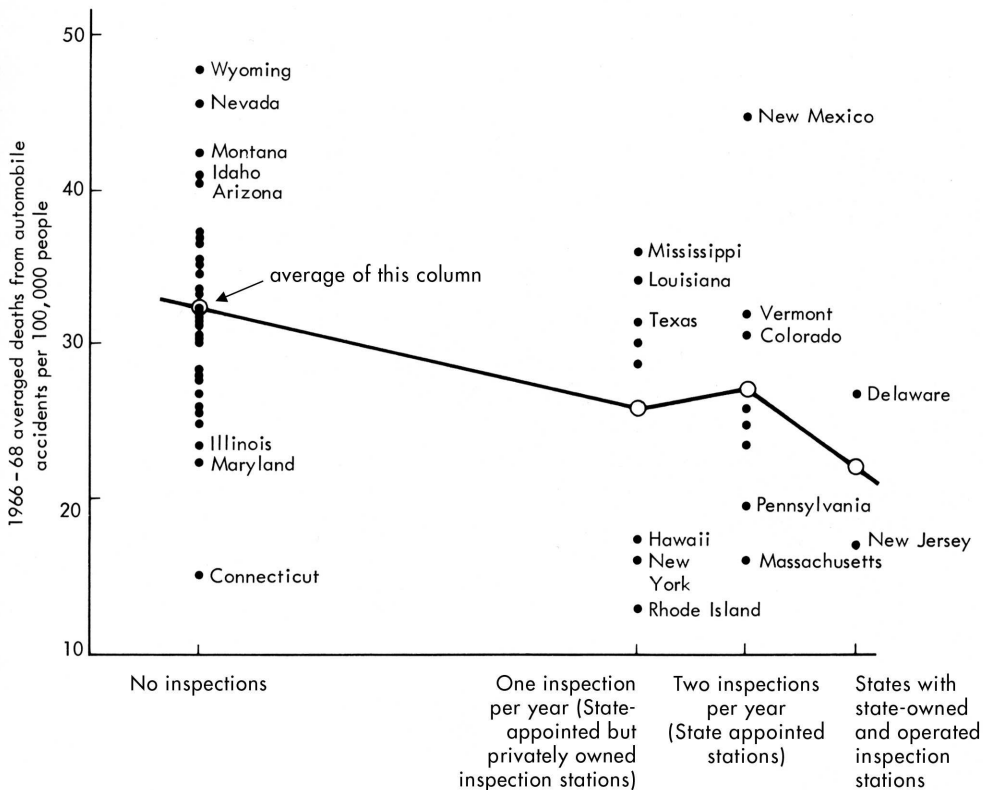
The measurement of the variables also raises questions. After discussing measurement difficulties, we will turn to the even more serious problem of the impact of other factors not now in the analysis.

1. The describing variable, inspections, is not measured particularly well. Right now, all the states are thrown into one of two mutually exclusive bins: either they have inspections or they do not. Such *dichotomous* or *dummy* variables, as they are called, should be used when there really are only two levels of the variable. In this case, since inspections differ widely in quality, a better way to assess the effects of inspections would be to classify states in several categories such as (a) no inspections at all, (b) relatively superficial inspections every other year, (c) superficial inspections every year, and (d) extensive inspections every year. If the fatality rate decreased as the quality of inspections improved, this would provide somewhat stronger support for the hypothesis that inspections do make a difference than the present evidence showing a difference only between inspected and uninspected states.

Figure 1-4 plots the average death rates for states without inspections and for states with three different “qualities” of inspections. There is a mild indication that the rate goes down as inspections improve, although the result is not striking.

2. States may differ in how they record deaths from auto accidents; such differences could, in turn, be linked to the presence or absence of inspections. In particular, states that have inspections might also have better investigation and reporting systems that distinguish traffic-accident deaths from, say, suicides and heart attacks that lead to motor-vehicle collisions. If there are such differences in recording deaths between the states, then in Figure 1-3 we would be observing

FIGURE 1-4 Death rate by inspection quality



a difference due to reporting of deaths rather than to inspections.

In such situations it is not enough to say: "There's error in the data and therefore the study must be terribly dubious." A good critic and data analyst must do more: he or she must also show *how* the error in the measurement or the analysis affects the inferences made on the basis of that data and analysis. Thus, in this case, two lines of argument are necessary to produce a legitimate statistical criticism. First, it is suggested that states may record deaths from automobile accidents differently. The second step is to suggest a mechanism by which such differences in the recording rate could lead to our present findings. Thus, it is further necessary to suggest not only that states differ in the way they record auto deaths, but also that these differences are related to whether a state has inspections. This seems to be a fair statement, since states with good procedures for analyzing the causes of death in automobile accidents might be those states with

activist state governments—indeed, the kind of state governments also likely to have a state inspection program.

3. The response variable is now measured in terms of *per capita* deaths—deaths per 100,000 people living in the state. But the individual driver might be more interested in the risk of death that is assumed for *each mile traveled* along the roads in that state. This reasoning suggests taking a look at the death rate *per hundred million miles* traveled, asking whether inspections reduce the risk of being killed for each mile driven. It turns out that in states with inspections, the death rate is 5.48 deaths per hundred million miles traveled, compared with 5.95 in states without inspections. The inspected states do somewhat better.

An interesting problem arises here in the computation of the mileage death rate. This rate is computed by taking the total number of deaths due to traffic accidents and dividing by the total number of miles traveled in the state. And how is the latter computed? Certainly the number of miles can't be measured directly. Rather, it is known how many gallons of gas are sold in each state, since all states have a gas tax yielding a few cents for each gallon of gas sold. The number of gallons sold are converted into number of miles traveled by assuming that cars get an average of about 12 miles for each gallon of gas. So, the overall computation is

$$\text{estimate of total miles traveled} = \frac{\text{revenue from gas tax}}{\text{gas tax rate (cents/gallon)}} \times 12 \text{ miles/gallon.}$$

For example, if the total tax revenue in a state was \$1,000,000 and the tax rate was \$0.10 per gallon, then 10,000,000 gallons were sold and an estimated 120,000,000 miles were traveled.¹¹ That is,

$$\frac{\$1,000,000}{\$0.10} \times 12 = 120,000,000.$$

4. Inspections cannot be expected to save all victims of auto accidents, simply because a large share of accidents are not the result of brake failure, bad tires, faulty steering, a missing tail light, or other mechanical defects detected and repaired as a consequence of inspections. A good many crashes are caused by factors that inspections

¹¹The actual calculation is somewhat more complicated, taking into account evaporation of gasoline, road differences between states, and so forth.

cannot remedy. For example, each year about 1500 people are killed in cars by trains at grade crossings. Probably another 500 die in the course of "hot pursuit" police chases.¹² An unknown (but probably significant) number of people choose the car as their suicide weapon. Finally, inspections will do little to reduce accidents due to drunken driving—and study after study clearly convicts drunken driving as the most important single factor leading to auto accidents. At least half of all fatal crashes involve a driver who had been drinking heavily and had a very high blood alcohol concentration at the time of the crash.

Thus some bias may enter the analysis because states may differ with respect to the proportion of accidents that can be prevented by inspections. Ideally, in the data analyst's heaven, the first step would be to determine the number of accidents *potentially* preventable by inspections and then, by comparing inspected and uninspected states, see whether inspections as currently used actually did prevent the accidents that they should have.

Discussing measurement of the quality of school facilities, Mosteller and Moynihan made the following observations, relevant to our discussion here, about "crude" versus "refined" measures in the study of policy:

. . . it is the experience of statisticians that when fairly "crude" measurements are refined, the change more often than not turns out to be small. Merely counting the number of laboratories in a school system is, in this sense, a "crude" measurement. It is possible to learn a good deal more about the quality of those laboratories. It could be that on further assessment the judgment to be had from the original crude measurement would be changed. But to repeat, statisticians would not leap too readily to that expectation. . . . Sadly, perhaps, in real life the similarities of basic categories are often far more powerful and important than the nice differences which can come to absorb individuals so disposed, but which really don't make a great difference in the aggregate.

The statistician would wholeheartedly say go ahead and make the better measurements, but he would often give a low probability to the prospect that the finer measures would produce information that would lead to different policy.

The reasons are several. One is that policy decisions are often rather insensitive to the measures—the same policy is often a good one across a great variety of measures. Secondly, the finer measures, as in the case of laboratories, can be thought of as something like

¹²This is a crude estimate; such estimates are obviously difficult to make accurately. See "500 Traffic Deaths Annually Attributed to Police 'Hot Pursuit,'" *The New York Times*, June 18, 1968.

weights. For example, perhaps one science laboratory is only half as good as another—well and good, let us count it 1/2. It turns out as an empirical fact that in a great variety of occasions, we get much the same policy decisions in spite of the weights. So there are some technical reasons for thinking that the finer measurement may not change the main thrust of one's policy. None of this is an argument against getting better information if it is needed, or against having reservations. More data cost money, and one has to decide where the good places are to put the next money acquired for investigations. If we think it matters a lot by all means let us measure it better.

Still another point about aggregative statistics is worth emphasizing for large social studies. Although the data may sometimes not be adequate for decisions about individual persons, they may well be adequate for deciding policy for groups. Thus we may not be able to predict which of two ways of teaching spelling will be preferable for a given child, but we may well be able to say that, on the average, a particular method does better. And then the policy is clear, at least until someone learns how to tell which children would do better under the differing methods.¹³

We have observed an association between inspections and lower death rates and have also considered some questions about that relationship. What do these results mean? Are there different explanations of the association between inspections and death rates?

Developing Explanations for the Observed Relationship

Many explanatory models begin by working with two variables: a response variable and a single describing variable. Usually, as the analysis develops, additional describing variables come into the model. Let Y denote the response (or dependent) variable and X the describing (or independent) variable. Begin by considering the notion that X causes Y :

$$X \longrightarrow Y.$$

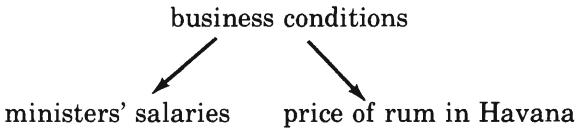
Returning to our example, we sought to find out whether

$$\begin{array}{ccc} \text{automobile} & & \text{low rate of traffic} \\ \text{inspections} & \longrightarrow & \text{fatalities.} \\ (X) & & (Y) \end{array}$$

¹³From "A Pathbreaking Report," in *On Equality of Educational Opportunity* by Frederick Mosteller and Daniel P. Moynihan, eds. Copyright © 1972 by Random House, Inc. Reprinted by permission of the publisher.

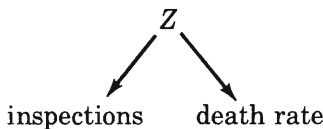
An observed association between two variables can occur for many reasons. There may be a causal relationship between the two variables. The relationship may occur simply by chance. Or X may *covary* with Y because both X and Y are jointly caused by some third factor Z . Thus, the observation that Y increases as X increases is consistent with many explanations.

Once we establish some kind of association between X and Y , the problem is what to make of it. There is supposedly a rather strong association over many years between the salaries of Presbyterian ministers and the price of rum in Havana—yet I doubt that we would want to suggest a causal relationship between the two. The apparent association between the ministers' salaries and the price of rum might arise because both were linked to some extent to the ups and downs of business conditions:



Thus, while salaries and rum prices apparently covary together with great regularity, it is not because ministers are spending their money for rum in Havana, but rather because both salaries and prices are linked to a common, third factor—the business climate. A correlation such as that between ministers' salaries and the price of rum is often called a *spurious correlation*; the relationship is spurious or misleading because the two variables are related only by some third cause.

Is there a possibility that the association between inspections and low death rates is spurious? Do states with both low rates and inspections have some third factor, Z , in common?



And how do we go about finding likely candidates for this variable Z ? Our substantive understanding of the problem may suggest some possible third variables to check for spurious correlation. In other cases, we simply might check through a number of possible variables that seem, for one reason or another, good possibilities. One useful guideline is simply to ask: What other factors are related to either X or Y ? In other words, are there any variables closely linked to

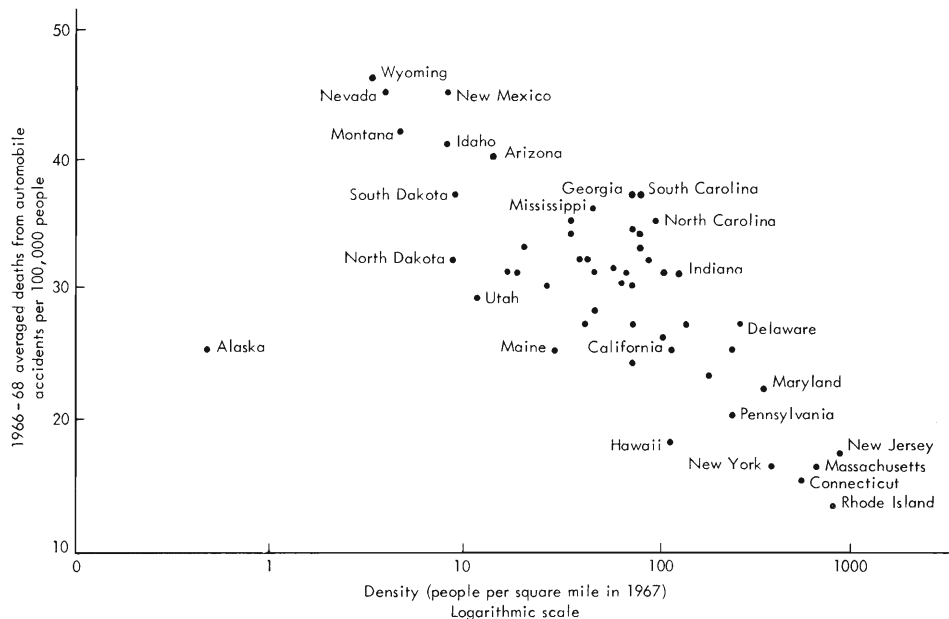


FIGURE 1-5 Death rate and density

death rate—and if so, are they also linked to the presence or absence of inspections?

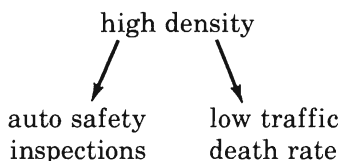
In looking for other such variables, we turn up several candidates: the density of the state, its weather conditions, and the percentage of young drivers. The density (number of people per square mile) is strongly related to the death rate: as the density of a state increases, the death rate decreases. In other words, thickly populated states such as Connecticut and New Jersey have low death rates from automobile accidents; thinly populated states such as Nevada and Wyoming have high rates. Figure 1-5 shows the relations between density and death rate for the fifty states.¹⁴ Such a pattern indicates a negative relationship, since the variables vary inversely; that is, as *X* gets bigger and bigger, *Y* tends to get smaller and smaller. States of *high* density, then, have *low* death rates; and states of *low* density have *high* rates. The scatterplot reveals a rather strong relationship between density and death rate, since the states progress in a relatively orderly fashion across the scatterplot.

Thinly populated states have higher fatality rates compared to thickly populated states because drivers go for longer distances at

¹⁴Density is plotted on a logarithmic scale in Figure 1-5 for reasons explained in Chapter 3. Alaska has been dropped from further analysis because of its atypical nature differing apparently from the other 49 states.

higher speeds in the less dense states. Accidents in states like Nevada and Arizona are probably typically more severe since they occur at a higher speed. It is not, however, just a matter of the number of miles driven, because there is also a fairly strong negative relationship between density and the *deaths per 100 million miles driven* in the state. Victims of accidents in the more thinly populated states, in addition to being involved in more severe accidents, are also less likely to be discovered and treated immediately, since both Good Samaritans and hospitals are more scattered in thinly populated states compared to the denser states.

Is the correlation between inspections of automobiles and low traffic death rates spurious? Given the strong relationship between density and the death rate, might there also be a relationship between density and the presence or absence of safety inspections? Are the high-density states (with their low death rates) more likely to have inspections? It looks that way; eight of the nine most thickly populated states have inspections, as compared with only one of the eight least dense states. This preliminary look suggests that the model



has some merit.

The density of a state's population certainly doesn't directly cause auto safety inspections. But a plausible argument explains the relationship between the two: the denser states tend to be the urbanized, industrialized, northeastern, politically competitive states with activist state governments—governments that would be more likely to inaugurate an inspection program. Looking at the data will help decide whether the relationship between inspections and reduced death rates is a spurious one resulting from the common element of density. To find out whether inspections have an effect, discounting the influence of density on the death rate, we will want to compare states at a similar level of density to see whether inspected states have lower death rate than uninspected states. To put it another way, it is necessary to hold density constant in order to observe the uncluttered (by density) effects of inspections on accident deaths.

Two different methods, *matching* and *adjustment*, help take into account the effects of density. Let us try it both ways here.

Matching simply involves taking states of roughly the same density

and seeing whether inspected states have lower rates than uninspected states within the density groupings. States are matched, then, with respect to density; often this procedure is called "controlling for" density. Table 1-1, comparing the average death rate for inspected and uninspected states for thinly, moderately, and thickly populated states, shows:

1. The averaged death rates are lowest in the thickly populated states and highest in the thinly populated states, regardless of whether they have inspections or not (in other words, the averaged rates decrease as we read across either the inspected-states row or the uninspected-states row).
2. At each level of density (thin, medium, and thick) the average death rate for the inspected states is lower than for uninspected states.

The average death rate for each of the six cells is computed by adding up the rates for the states in given cell and dividing by the number of states in the cell. This average or mean rate is very sensitive to extreme values; for example, for the thinly populated states with inspections, the three states have the rates 28.8, 30.9, and 45.0. New Mexico, at 45.0, forces the average up to almost 35, even though two of the three states are actually close to 30.

The division and assignment of states into three categories is perfectly arbitrary. Many other divisions are probably just as good. Table 1-2 shows a slightly different set of categories; it differs somewhat from Table 1-1 because the shuffling of a few states from one category to another affects the averages to some extent. Table 1-2, like Table 1-1, however, shows that some relationship remains between inspections and a reduced death rate even when the effects of density are controlled.

The matching procedure often helps inform the reader what is going on in the data: Tables 1-1 and 1-2 clearly display the effect of inspections at the three density levels and also the effect of density at each inspection level. Matching has some defects, chiefly that it is difficult to do a very good job of matching in complex situations without a large number of cases. In Table 1-1 we have not really matched the states in a very satisfactory way by throwing them into three bins labeled "thin," "moderate," and "thick." A good deal of variation still remains within each of the three levels of density. For instance, both Wyoming (density = 3.2 people per square mile) and Oregon (density = 20.8) are described as "thinly populated," although they differ widely in density. Thus by putting the states into only three categories we lose some information about one of the key variables (density). Before

TABLE 1-1
Inspections, Density, and Average Death Rates

	Density					
	Thin		Medium		Thick	
	Average	N	Average	N	Average	N
States without inspections	38.5	9	31.5	16	23.6	6
States with inspections	34.9	3	28.4	9	18.3	6

Definitions:

- Thin = density less than or equal to 25 people per square mile.
 Medium = more than 25 and less than 125 people per square mile.
 Thick = 125 or more people per square mile.
 Average = mean death rate for states in that category (computed by adding up the death rates for all the states in that category and dividing by the number of states in that category).
 N = number of states in that category.
 Total = 49 states (all states except Alaska).

ORIGINAL DATA (STATES AND THEIR DEATH RATES)

Density

		Thin		Medium		Thick	
<i>States without inspections</i>	Arizona	38.8	Alabama	31.2	Connecticut	14.7	
	Idaho	40.2	Arkansas	34.2	Illinois	23.0	
	Montana	42.2	California	25.4	Indiana	31.1	
	Nebraska	30.5	Florida	31.4	Maryland	22.0	
	Nevada	45.4	Georgia	36.9	Michigan	26.5	
	North Dakota	31.7	Iowa	31.3	Ohio	24.5	
	Oregon	33.3	Kansas	30.0			
	South Dakota	37.0	Kentucky	33.0			
	Wyoming	48.0	Minnesota	27.7			
			Missouri	30.5			
			North Carolina	35.1			
			Oklahoma	35.0			
			South Carolina	36.6			
			Tennessee	31.5			
			Washington	28.1			
		Wisconsin	27.4				
<i>States with inspections</i>	Colorado	30.9	Louisiana	34.1	Delaware	27.0	
	New Mexico	45.0	Maine	24.7	Massachusetts	16.4	
	Utah	28.8	Mississippi	36.0	New Jersey	17.3	
			New Hampshire	23.8	New York	16.1	
			Texas	31.5	Pennsylvania	19.8	
			Vermont	32.0	Rhode Island	13.0	
			West Virginia	30.1			
			Virginia	26.0			
			Hawaii	17.8			

TABLE 1-2
 Inspections, Density (Different Division), and Average Death Rates

	Thin		Medium		Thick	
	Average	N	Average	N	Average	N
States without inspections	37.6	11	32.1	11	26.0	9
States with inspections	32.4	4	31.2	6	19.2	8

Identical to Table 1-1 except:

- Thin = density less or equal to than 37 people per square mile.
 Medium = more than 37 and less than 100 people per square mile.
 Thick = 100 or more people per square mile.

classifying both Wyoming and Oregon as thinly populated, we knew that they differed by such-and-such amount in their densities. But now, in Table 1-1, this information is not used in the analysis, and the two states are treated as if they were alike. The situation is just as troubling for the states in the thickly populated category. Here, the states range from a density of 138.3 people per square mile in Indiana up to New Jersey with 929.8.

One limitation of matching, then, is that quite often the match is not very accurate. A second limitation is that if we want to control for more than one variable using matching procedures, the tables begin to have combinations of categories without any cases at all in them, and they become somewhat more difficult for the reader to understand. For example, if states were matched with respect to density (three categories in this case) and, in addition, their weather (say five categories), the fifty states would be scattered over fifteen different combinations of density and weather conditions (and some combinations might not even exist empirically—for example, a warm, dry state that was also densely populated). When the inspection classification was added, the fifty states would then be classified into thirty categories. The scattering of cases over many different cells (or combinations of different levels of variables) of the table can be avoided by collapsing categories (using, say, two levels of density instead of three)—but then, of course, states become less and less well matched, and the effects of density are less well controlled because of the wide variations in density in supposedly “matched” groups.

Adjustment, the other procedure for controlling the effects of a third variable, sometimes partially overcomes these difficulties. By standardizing the death rate of each state for the density of that

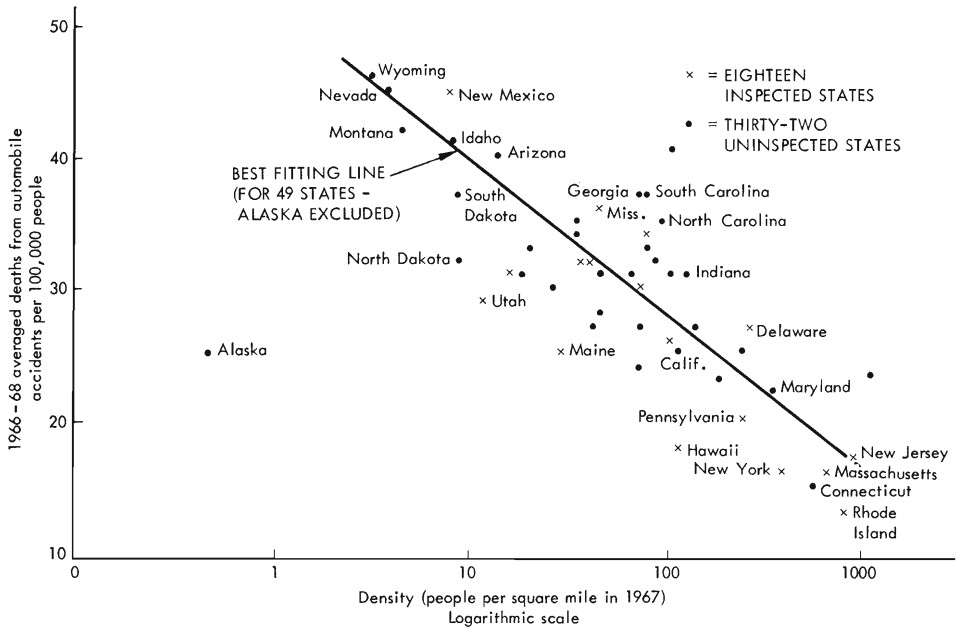


FIGURE 1-6 Fitted line: death rate and density

state, the adjustment procedure takes out the effect of density on the death rate, producing what might be called a “density-standardized death rate.” We can employ the procedure informally merely by looking at the scatterplot (Figure 1-6), which shows the plot of the death rate against density for the inspected and uninspected states. The line fitted to the points here is the line that best fits the relationship between density and deaths.

The line makes what is essentially an average prediction: given that a state has a certain density, the line predicts that state’s death rate. Some states lie below the prediction line, indicating that they have a lower death rate than predicted by their density. States that lie above the line have a higher death rate than predicted. If inspected states have a lower death rate—for their density level—than uninspected states, then they should tend to lie below the line and below the points representing the uninspected states in the same region of density on the scatterplot. In other words, the little crosses (representing the inspected states) should, at a given density level, tend to lie below the dots (representing the uninspected states) if inspections have an effect after controlling for density. Although no vivid effect appears in Figure 1-6, it is possible to see a slight tendency indicating lower rates in inspected states.

Let us now formalize the adjustment procedure and take out the

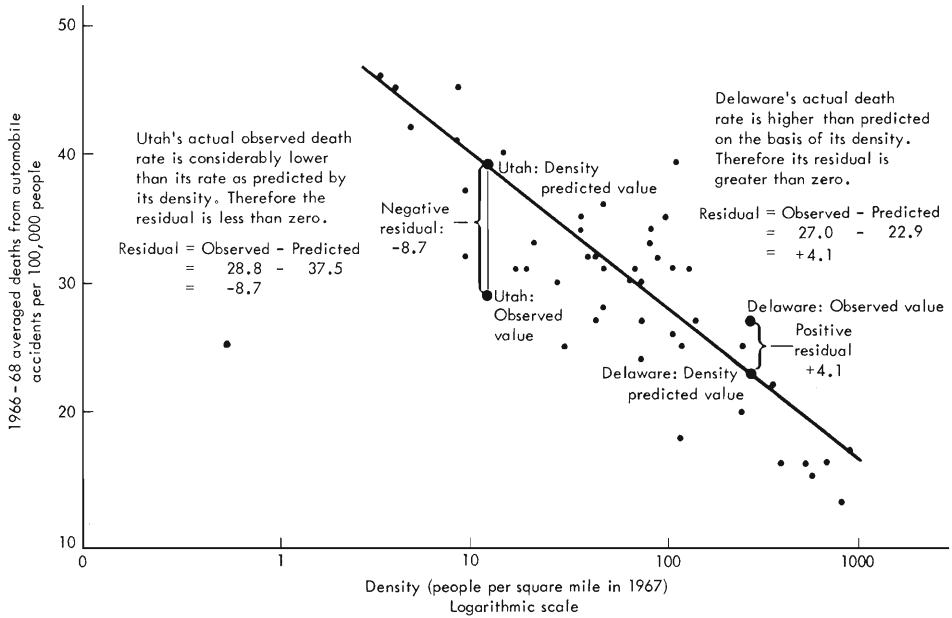


FIGURE 1-7 Residuals from fitted line

effects of density mathematically. The line fitted to the points represents the predicted death rate for a given density. Thus for each state there is a predicted death rate—a prediction based on its density. Also, we know the *actual* death rate in each state. The difference between the actual, observed death rate for a state and the predicted death rate represents that part of the death rate that is unaccounted for by the state's density. The difference between the observed and the predicted death rate is called the *residual*:

residual	=	actual observed	-	predicted (by
for a		death rate		density) death rate
given state		for that state		for that state

Thus the residuals for all the states are computed simply by subtracting the density-predicted death rate from the actual rate.¹⁵ Each residual can be viewed as a death rate adjusted for density; it is that part of the death rate that is unexplained by density. Figure 1-7 shows the logic. Generating a predicted death on the basis of density and then examining the residual death rate is, in effect, a statistical

¹⁵The computational method is described in Chapter 3.

way of matching or equating all states with respect to density. The examination of residuals is a powerful tool for the analysis of data, since the residual represents that part of the variation in the response variable that remains unexplained after looking at a set of describing variables. The residuals measure what remains to be explained in the response variable. New explanations can be developed by seeing how the residuals are related to other describing variables. Examples and further details are found in Chapters 3 and 4.

Figure 1-8 shows the residuals (or the density-adjusted death rates) for the inspected and the uninspected states. Generally, those states with a lower death rate than expected are those states that have inspections—with Mississippi, Louisiana, and New Mexico being very prominent exceptions. On the average, states with inspections have a rate 1.63 deaths per 100,000 people *lower* than expected, and states without inspections have a rate of 0.90 deaths per 100,000 population *higher* than expected—yielding a difference of 2.5 deaths per 100,000 between inspected and uninspected states after adjustment of the rates for density. While the difference is neither large nor sure, it does favor inspections. The difference might suggest that if inspections were implemented by all states, perhaps an additional 2500 lives would be saved each year. This is far from certain. Greater certainty might be obtained by taking other variables into account. But to increase substantially the credibility of the view that inspections make a difference would require a well-designed experiment. The nonexperimental data examined here can only give small hints about what is going on.

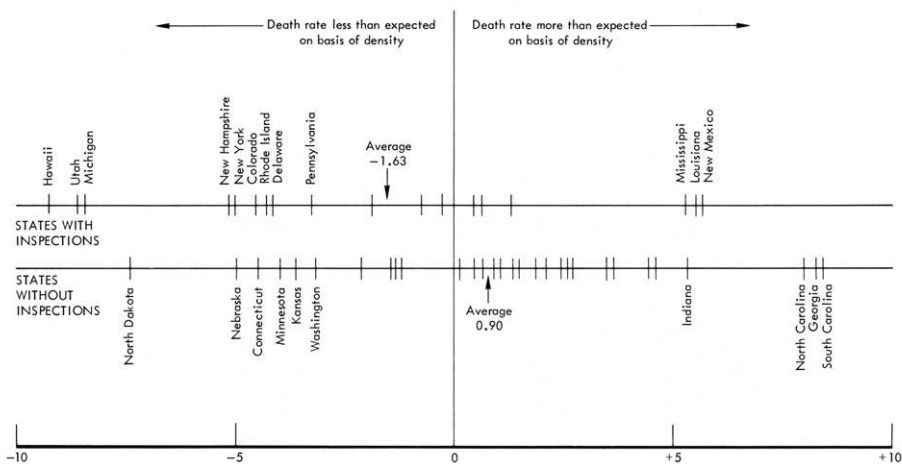


FIGURE 1-8 Residuals—the density-adjusted death rate—for inspected and uninspected states

Compulsory inspections of automobiles, by getting some mechanical defects straightened out, might produce a modest reduction in the death rate from car crashes. If intervention at the level of the car owner has effects of the size observed in this study, then what additional measures, beyond inspections, might cut the death and injury rate from automobile accidents? As mentioned earlier, efforts to reduce drunken driving may be helpful. But safety efforts at the level of the individual driver are limited; as Moynihan wrote:

There is not much evidence that the number of accidents can be substantially reduced simply by altering the behavior of drivers while maintaining a near universal driving population. It may be this can be done, but it has not been done. This leads to the basic strategy of crash injury protection: it is assumed that a great many automobile accidents will continue to occur. That being the case, the most efficient way to minimize the overall cost of accidents is to design the interior of the vehicles so that the *injuries* that follow the *accidents* are relatively mild. An attraction of this approach is that it could be put into effect by changing the behavior of a tiny population—the forty or fifty executives who run the automobile industry.¹⁶

¹⁶Moynihan, *op. cit.*, p. 12.

Costs and Unquantifiable Aspects

To conclude let us briefly consider some of the costs of inspections and look at some aspects of the problem that are not quantifiable.

Inspections, as noted earlier, have significant costs. Almost all of these costs, direct and indirect, fall on the individual car owner. Inspections, therefore, produce few pressures on or incentives for automobile manufacturers to build safer cars free of mechanical defects. Under an inspection system, if the headlights of a car are misaligned in the factory or if a tail light burns out, the car owner pays the cost of fixing the defect when it is discovered in the inspection. Not only is there no cost to the manufacturer for having produced a car with a defect, but indeed there is a further profit to be made on the replacement part correcting the defect. Thus inspections are a limited strategy for coping with car crashes because of their modest effects, their significant costs, and their failure—if it may be called that—to snowball into further safety efforts.

Earlier, some crude estimates of the economic costs of inspections were given—the figure running to perhaps \$500 million in those states with inspections. There are also political and social costs of programs (such as inspections) which require coercion by the threat of arrest and fine of large number of citizens (in this case, 80 million car owners). While the total experiences of most citizens with their government occur in similarly coercive and bureaucratic contexts—such as the income tax, the draft, traffic tickets, and auto licensing—what are, in fact, the long-run costs of bureaucratic and arbitrary impingements upon citizens by the government? Do some citizens consequently become alienated and cynical about its performance? Does the modest coercion involved in inspection programs lead to the eventual acceptance of increasingly more severe coercion?

Since it is difficult to measure certain kinds of political and social costs, as well as benefits, of a program, such unmeasurable factors sometimes receive less emphasis than they should. (On the other hand, bizarre estimates of such costs may go unchallenged for the lack of data to prove them wrong.) For example, in the judicial process, it is easy to measure police performance in terms of the numbers of arrests made; but it is more difficult to assess performance with respect to equal or fair treatment. Or, to take another example, the

apparently huge costs of smoking cigarettes—the years of life lost to early death, the excess illness among smokers, the fires started by smoking—have been measured carefully and extensively in the last twenty years. In contrast, the gratification received from smoking by the smoker cannot be ascertained; and presumably such information has at least modest relevance to decisions about public policy toward smoking.

Our inability to measure important factors does not mean either that we should sweep those factors under the rug or that we should give them all the weight in a decision. Some important factors in some problems can be assessed quantitatively. And even though thoughtful and imaginative efforts have sometimes turned the “unmeasurable” into a useful number, some important factors are simply not measurable. As always, every bit of the investigator’s ingenuity and good judgment must be brought into play. And, whatever unknowns may remain, the analysis of quantitative data nonetheless can help us learn something about the world—even if it is not the whole story.

CHAPTER 2

Predictions and Projections: Some Issues of Research Design

"There will be no nuclear war within the next fifty years."

"In the period 1965–70, Mao Tse-tung and De Gaulle will die."

"Major fighting in Viet-Nam will peter out about 1967; and most objective observers will regard it as a substantial American victory."

"In the United States Lyndon Johnson will have been re-elected in 1968."

—Ithiel de Sola Pool¹

Introduction

Projections of the future can be useful or embarrassing, depending on their accuracy. The assumption that a wide range of factors remain constant or continue to change at current rates can quickly crumble.² And yet how imbedded in our thought is the idea that the future is a straightforward projection of the past: we may doubt the optimism of Professor Pool's first prediction if only because of the failure of the other predictions on the list. At least, unlike some predictions, these have the modest virtue of being explicit, and it is easy to tell whether they went wrong.³

¹"The International System in the Next Half Century," in Daniel Bell, ed., *Toward the Year 2000: Work in Progress* (Boston: Beacon Press, 1967), pp. 319–20.

²A very useful discussion of the assumptions behind many projections is Otis Dudley Duncan, "Social Forecasting—The State of the Art," *The Public Interest*, no. 17 (Fall 1969), 88–118.

³On previous prophecies, see Arthur M. Schlesinger, "Casting the National Horoscope," *Proceedings of the American Antiquarian Society*, 55 (1945), 53–93.

Almost all efforts at data analysis seek, at some point, to generalize the results and extend the reach of the conclusions beyond a particular set of data. The inferential leap may be from past experiences to future ones, from a sample of a population to the whole population, or from a narrow range of a variable to a wider range. The real difficulty is in deciding when the extrapolation beyond the range

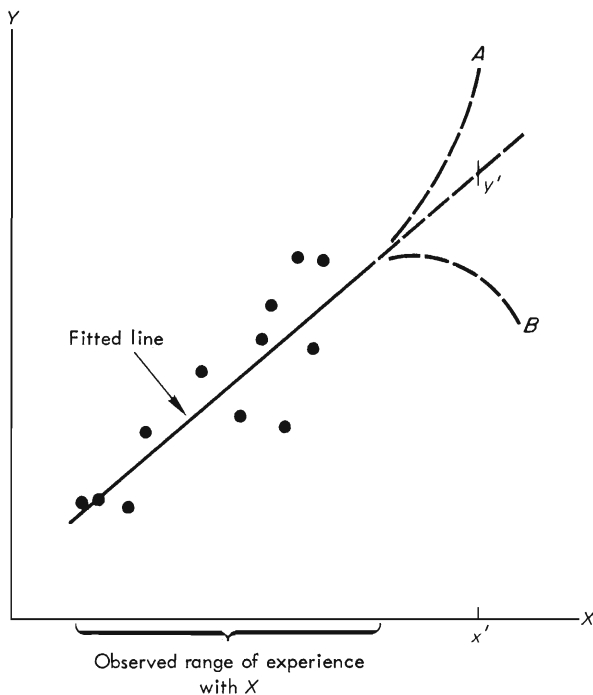


FIGURE 2-1 Problem of simple extrapolation

Q: Should the fitted line be extended to predict the value y' for the new observation x' (which is outside the range of previous experience with the x -variable)? Or, is A or B a better model?

A: "A priori nonstatistical considerations . . ."

of the variables is warranted and when it is merely naive. As usual, it is largely a matter of substantive judgment—or, as it is sometimes more delicately put, a matter of "a priori nonstatistical considerations" (Figure 2-1).

If the observed variation in a variable is small relative to its total possible variation, then the extension of the inference based on a narrow range of observations is less warranted than extrapolation

based on a wider range of observed variations. Equally obvious is the observation that the risk of error is less if the extrapolated value is "close" to the previous pattern of experience rather than greatly different, other things being equal. In some cases it may be useful to conduct trial runs at extrapolation by using a fraction of the available data to produce a fitted curve, using the remaining data to test the accuracy of the extrapolated results. Obviously if the conditions governing a relationship change in relevant respects, the effort at extension of results is in danger of making errors.

Simple extrapolation involves the extension of results outside the range of experience of a single describing variable. A more subtle situation arises in the multivariate case involving extrapolation beyond the range of the *combination* of experience jointly observed in two or more describing variables. Karl A. Fox has described this situation as "hidden extrapolation."⁴

Figure 2-2 shows the pattern of correlation between two describing variables. Assume these two describing variables, X_1 and X_2 , are used in combination to predict a response variable, Y . The situation appears to be relatively satisfactory because there is a wide range of experience with both X_1 and X_2 . But note how little experience there is concerning certain *combinations* of X_1 and X_2 —since all the points representing joint occurrences of X_1 and X_2 are contained in the narrow band surrounding the line. There is no experience with combinations such as low X_1 -high X_2 (in the upper left of the rectangle) or high X_1 -low X_2 (lower right) and how such unobserved combinations of X_1 and X_2 might affect the response variable. The response variable may behave very differently for such combinations of X_1 and X_2 . Thus a prediction equation, predicting Y from X_1 and X_2 , may be quite misleading if applied to situations in which X_1 and X_2 occur in combinations different from those observed here.

Thus the extension of the inference over all combinations of X_1 and X_2 may founder on the possibility of an interaction effect between X_1 and X_2 in their influence on Y in the region of the combinations with which there is no experience. The problem arises because of limited experience with the *joint* relationship of X_1 and X_2 , even though there may be extensive experience with the entire range of each variable taken singly. Thus the name, "hidden extrapolation."

The problem arises in any predictive study involving correlated describing variables. Figure 2-3 shows the narrowed range of joint experience in the case of three correlated describing variables.

We diagnose the problem by considering the scatterplots of the

⁴This discussion is based on Karl A. Fox, *Intermediate Economic Statistics* (New York: Wiley, 1968), pp. 265-66.

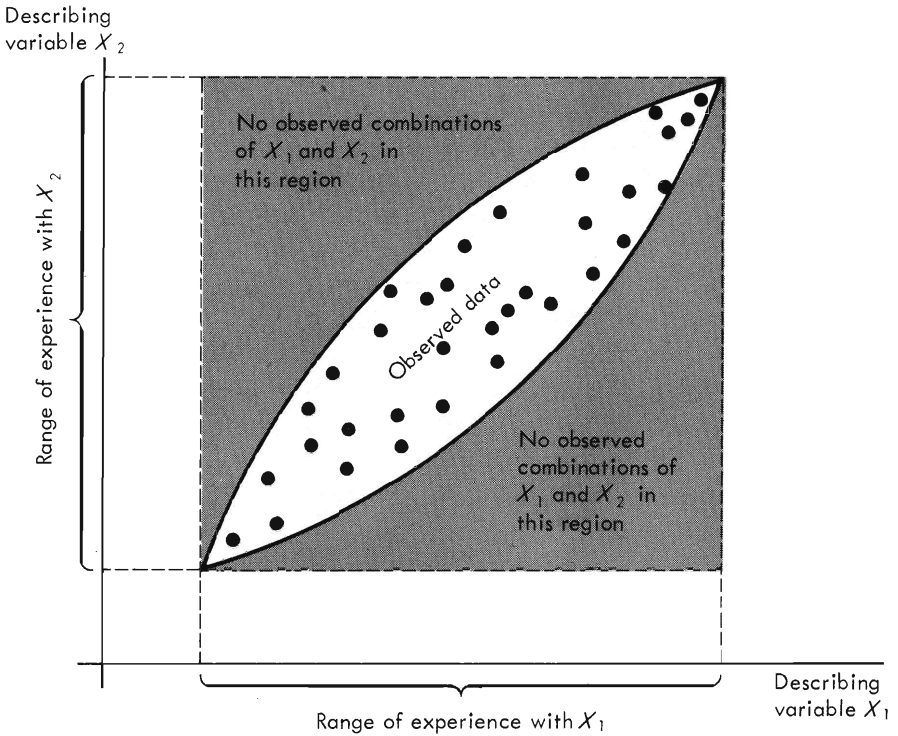


FIGURE 2-2 Correlation between two describing variables

relationships between the describing variables and by looking over the original joint observations. Cures for the difficulty include the collection of additional data, particularly of "deviant cases" in areas outside the previously experienced combinations of describing variables.

Let us now turn to several examples illustrating and evaluating methods of prediction. These case studies show different statistical tools in action. Note, however, that the central consideration in most cases is the research design, rather than the mechanics of using the statistical tool. Mosteller and Bush make this point quite sharply:

We first wish to emphasize that formal statistics provides the investigator with tools useful in conducting thoughtful research; these tools are not a substitute for either thinking or working. A major goal for the statistical training of students should be statistical thinking rather than statistical formulas, by which we mean specifically: thinking about (1) the conception and design of the study and what it is that is to be measured and why, (2) the definitions of the terms being used, and how modifications in definition might change both the outcome and the interpretation of a study, (3) sources of variation in every part of the study, including such things as

individual differences, group and race differences, environmental differences, instrumental or measuring errors, and intrinsic variation fundamental to the process under investigation. In no circumstances do we think that sophisticated analytical devices should replace clean design and careful execution, unless very unusual economic considerations arise. However, it may be worth remarking that crude data collected as best the investigator could may require the most advanced statistical tools. Here a quotation from Wallis may be appropriate:

In general, if a statistical investigation . . . is well planned and the data properly collected the interpretation will pretty well take care of itself. So-called "high-powered," "refined," or "elaborate" statistical techniques are generally called for when the data are crude and inadequate—exactly the opposite, if I may be permitted an obiter dictum, of what crude and inadequate statisticians usually think."⁵

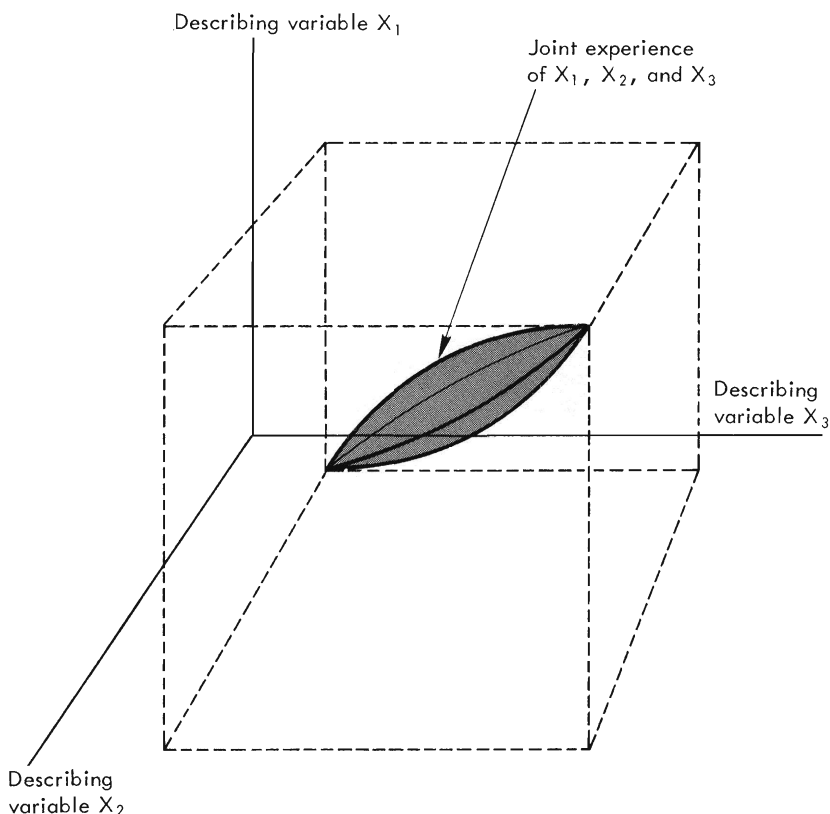


FIGURE 2-3 Range of joint experience—three describing variables

⁵Frederick Mosteller and Robert R. Bush, "Selected Quantitative Techniques," in Gardner Lindzey, ed., *Handbook of Social Psychology: Vol. I, Theory and Method* (Cambridge, Mass.: Addison-Wesley, 1954), p. 331. The passage by Wallis is found in W. Allen Wallis, "Statistics of the Kinsey Report," *Journal of the American Statistical Association*, 44 (1949), p. 471.

Problem in Prediction: The National Crime Test and a Cancer Test

Assessing the quality of a prediction or extrapolation can sometimes be a tricky matter. Consider the following example, which reveals the interplay between the properties of the predictive device and the tested population.

A proposal was once made that every 6-, 7-, and 8-year-old child (a total of 13 million in all) be given psychological tests to identify potential "criminality" in order that the supposed lawbreakers of the future be given some sort of treatment. The proposal encountered a storm of moral, legal, and technical criticism which led to its apparent abandonment. One of the technical flaws, which also serves to emphasize the moral and legal criticism of the proposal, is shown in the following model. Assume the National Crime Test has the following hypothetical properties:

1. It will successfully identify 40 percent of those arrested in the future. (Unfortunately, a child's "identification" by the NCT might help insure his future arrest through the mechanism of a self-fulfilling prophecy, operating with respect to the child or the police or both. Perhaps even NCT scores would be used to convince a jury of the guilt of the accused—thereby further increasing the "accuracy" of the prediction.)
2. It will also correctly classify 90 percent of those children who will not be arrested in the future.

Do these characteristics of our hypothetical NCT indicate it is a useful predictor of criminality? It might seem so, since it does identify four out of ten of the future "bad guys" and nine out of ten of the "good guys." But let us look into the errors in prediction made by a test with these characteristics. Assuming that three percent of these children will, later in life, commit a serious crime, we can construct Table 2-1, which shows the predictive performance of the NCT.

The table shows the errors made in the test; let us consider the "false positives" in which the test predicts criminality incorrectly. The upper righthand corner of the table shows 1,261,000 false positives compared to 156,000 correct predictions of criminality. Thus for every correct prediction of future difficulties, there are eight incorrect ones! In this light, such a test would be unacceptable to most people—even though its predictive characteristics, as originally expressed, seemed impressive. Furthermore, the assumptions we made about the predictive powers of such tests were, if anything, much too generous, given the poor performance of psychological tests of "criminality."

TABLE 2-1
Hypothetical (Fortunately) National Crime Test

		Reality	
		<i>Criminal</i>	<i>Noncriminal</i>
Test predicts	<i>Criminal</i>	156,000	1,261,000
	<i>Noncriminal</i>	234,000	11,349,000
		<u>390,000</u>	<u>12,610,000</u>
Total = 13,000,000			

COMPUTATIONS:

3 percent of 13,000,000 children will commit a serious crime:

$(.03)(13,000,000) = 390,000$ children. NCT accurately predicts 40 percent:

$(.40)(390,000) = \underline{156,000}$

97 percent of 13,000,000 are not future criminals:

$(.97)(13,000,000) = \underline{12,610,000}$. NCT accurately predicts 90 percent:

$(.90)(12,610,000) = \underline{11,349,000}$.

Consider another example of the same problem. A hypothetical test for cancer has the following characteristics:

1. $\Pr(\text{test positive} \mid \text{cancer}) = .95$. This conditional probability indicates that the test reads "positive" 95 percent of the time given that the person tested in fact has cancer.
2. $\Pr(\text{test negative} \mid \text{no cancer}) = .96$.

In other words, the test correctly identifies, on the average, 95 out of 100 of those who do have cancer and also 96 out of 100 of those who do not have cancer. These characteristics give the following table of probabilities:

		Reality	
		<i>Cancer</i>	<i>No cancer</i>
Test predicts	<i>Positive</i>	.95	.04
	<i>Negative</i>	.05	.96
		<u>1.00</u>	<u>1.00</u>

Now assume that one percent of those tested actually do have cancer; that is, $\Pr(\text{cancer}) = .01$. (This is an unconditional probability, since it depends upon no given prior condition.) Note that since only one percent of those tested have cancer, the flow of those tested is mainly down the righthand column of the table of probabilities.

What proportion of false positives (and false negatives) will be

TABLE 2-2
Computation of Probabilities

We have the following data:

$$\Pr(\text{cancer}) = .01$$

$$\text{Therefore } \Pr(\text{not cancer}) = 1.00 - .01 = .99.$$

Similarly,

$$\Pr(\text{test positive} \mid \text{cancer}) = .95, \text{ and therefore}$$

$$\Pr(\text{test negative} \mid \text{cancer}) = .05.$$

Also,

$$\Pr(\text{test negative} \mid \text{no cancer}) = .96, \text{ and therefore}$$

$$\Pr(\text{test positive} \mid \text{no cancer}) = .04.$$

The problem is to compute $\Pr(\text{cancer} \mid \text{test positive})$, which equals, by Bayes' theorem:

$$\frac{\Pr(\text{test positive} \mid \text{cancer}) \Pr(\text{cancer})}{\Pr(\text{test positive} \mid \text{cancer}) \Pr(\text{cancer}) + \Pr(\text{test positive} \mid \text{not cancer}) \Pr(\text{not cancer})} = \frac{(.95)(.01)}{(.95)(.01) + (.04)(.96)} = .19.$$

produced by the test? One way to answer with respect to false positives is to compute $\Pr(\text{cancer} \mid \text{test positive})$ —the probability that a person has cancer, given that the test reads positive. This can be done, using the appropriate equations for conditional probabilities, shown in Table 2-2. Another way to handle the problem is to consider what happens when, say, 10,000 people are screened for cancer using the hypothetical test. Computations analogous to those in Table 2-1 yield the following expected results:

		Reality	
		Cancer	No cancer
Test predicts	Positive	95	396
	Negative	5	9,504

and therefore

$$\Pr(\text{cancer} \mid \text{positive}) = \frac{95}{95 + 396} = .19.$$

Thus about 19 percent of those indicated positive will actually have cancer; 81 percent of the positives will be false. The decision whether this is a good test depends upon the cost of such false positives and their consequent detection as well as the benefits that derive from

the detection of the disease. Perhaps such a test would be most useful as a screening device to indicate patients needing further tests.

Similar arguments apply to the use of lie detectors, the prediction of juvenile delinquency on the basis of family background, and the use of "preventive detention."⁶ The reason the original qualities of the prediction seem to collapse when the test is applied to data is that, in these two cases, the quality to be detected is rather rare. Therefore, even though the hypothetical cancer test correctly predicts cancer 95 percent of the time and noncancer 96 percent of the time, so many people (99 percent in our example) flow through the right (noncancer) side of the table of probabilities that even the low error rate (4 percent) produces a large number of errors relative to the number of correct predictions of cancer. If, on the other hand, *half* the tested population had cancer, then the expected table (for 10,000 people) would be:

		Reality	
		<i>Cancer</i>	<i>No Cancer</i>
Test predicts	<i>Positive</i>	4750	200
	<i>Negative</i>	250	4800

This is pretty sensational predicting!

The *properties of the test are the same* in both cases, but the populations tested differ with respect to the distribution of the characteristic to be detected. Thus a test which does a good job of prediction on one population may not perform so well on a second trial if distribution of the characteristic sought differs markedly in the second population. Thus it will be worthwhile to try out—if only by working through the arithmetic as we have done here—the test on a population for which the distribution of the characteristic to be predicted is the same as the population for which the ultimate prediction is to be made. Note that the two numbers $\text{Pr}(\text{positive} \mid \text{cancer})$ and $\text{Pr}(\text{negative} \mid \text{not cancer})$ were not enough to describe adequately the performance of the prediction. Instead, a third piece of information, in this case $\text{Pr}(\text{cancer})$, was necessary to permit an adequate assessment of the performance of the test for that population.

⁶See Jerome H. Skolnick, "Scientific Theory and Scientific Evidence: An Analysis of Lie-Detection," *Yale Law Journal*, 70 (April 1961), 694-728; and Travis Hirschi and Hanan C. Selvin, *Delinquency Research* (New York: Free Press, 1967), chap. 14.

Finally, some very high rates of successful “prediction” should not fool us. After all, we can achieve 99 percent “accuracy” simply by predicting that no person has cancer. Since 99 percent of the people in our example don’t have cancer, the rule is 99 percent “accurate” in a sense, although next to worthless medically.

Election-Night Forecasting

Each election night, when the polls have closed and the votes are being counted, the three television networks forecast the electoral outcome on the basis of early, partial returns—often needing only a few percent of the vote to predict accurately the final outcome. The networks invest millions of dollars in their electoral coverage, which allows their viewers to learn the results of the election several hours earlier than they might otherwise. Although this is perhaps a small yield for the investment, the scramble for early returns needed for the projection of the winner might, in some places in some elections, discourage corrupt election officials from greatly altering the real count of the vote—since the pressure of getting the vote count in may reduce the time needed to fix the returns.

For example, pressures for a timely count may curb such abuses as those in Illinois in the 1968 tabulation:

For days before the election, the Chicago papers were full of tales of heavy crops of bums and derelicts being registered in West Side flophouses to provide the names for a fine Democratic turnout. And suspicion became certainty in the press rooms . . . when it was learned that “computer breakdowns” and “disputed vote counts” were holding the Illinois decision back. Veteran reporters could be heard explaining . . . how the game was played in Illinois: how both the iron Mayor and his Republican enemies downstate would “hold back” hundreds of precincts in an effort to finesse each other to give a hint of the size of the total they had to beat; how they would release a few precincts as bait to lure the other man into giving away some of his. . . .⁷

This suggests that the count of the vote is a rather unusual statistic. For most social and economic indicators, there is a tradeoff between timeliness and accuracy: the quicker we get the information, the greater the error. Sometimes the making of economic policy has been based on very short-run economic statistics—with a resulting reliance

⁷Lewis Chester, Godfrey Hodgson, and Bruce Page, *An American Melodrama: The Presidential Campaign of 1968* (New York: Viking, 1969), pp. 760–61.

on less accurate statistics—and more accurate figures might well have produced a different policy. In contrast to the usual case, however, a slow count of the vote often indicates vote fraud, or at least the opportunity for vote fraud.⁸

Although they may, in passing, reduce vote fraud, the central concern of the networks is to forecast the winner of the election (and, secondarily, the winner's share of the vote) on the basis of scattered and very incomplete returns. Two methods, both interpreting early returns with reference to a historical baseline drawn from previous elections, have been favored: (1) comparison of tonight's returns with the returns from previous elections at the same stage of the count and (2) comparison of tonight's returns from various counties with the returns from previous elections from those same counties.

The first method begins by constructing, on the basis of a previous election, a curve showing the relationship between the proportion of the vote reported and the proportion of the reported vote for the Democratic (or Republican) candidate. Figure 2-4 shows one such pattern, indicating that in this case a Democratic candidate who has more than about 40 percent of the vote when less than about 70 percent of the vote has reported can expect to win rather easily when all the returns are in. Such a pattern might result from the early reporting of certain Republican areas and a slower count in heavily Democratic areas. Thus the curve—called a "mu curve"—helps adjust for the bias favoring one party or the other in the sequence of early returns. Figure 2-5 indicates how this might be done. Tonight's returns are compared with the historical pattern of reporting, an appropriate adjustment for reporting bias is made, and the final projection is put on the air. In practice, the method is fancied up a bit—but still its basic defect persists: it relies on the assumption that the order in which the vote is reported remains the same from election to election. This assumption has led to several predictive disasters, and now mu curves only supplement other, more solidly based techniques.

One such predictive botch occurred during an election when a heavily Republican state first introduced voting machines. As a result, that state's flood of Republican ballots came in hours earlier than usual; the mu curve, believing that these were the same votes it saw in each election every four years, quickly projected a Republican landslide for president. Hours and hours later, John Kennedy won one of the closest presidential contests in history.

⁸The problem of inaccurate counts of the vote is not unimportant; political observers guess that two or three million votes are stolen, miscounted, or changed in a U.S. presidential election. Nobody has a good guess about the partisan advantage, if any, resulting from stolen votes. The advantage differs by state.

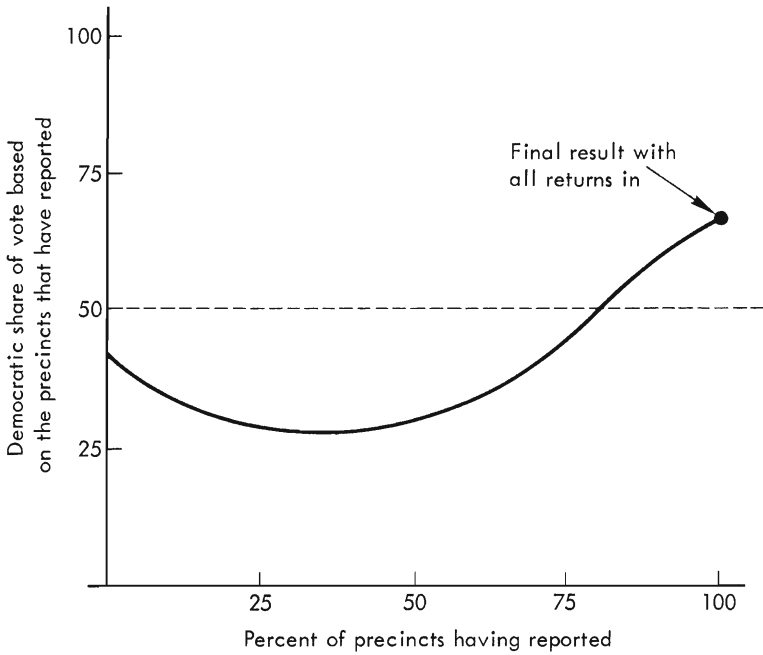


FIGURE 2-4 Historical pattern of the vote as more and more precincts report their returns on election night

Some practitioners patch up their mu curves by taking into account expected changes in the order of reporting:

In deriving mu curves which are empirical in nature—they have to be—one must take into very careful consideration whether or not there have been any changes in voting patterns resulting from voting machines, or changes in poll closing times. Where there are such changes—and in every election we find that there are some—the mu curves have to be suitably adjusted in order to render them suitable.⁹

This sort of repair requires knowledge *in advance* of those changes in election procedures that might affect the sequence of the vote report—and must then guess how much earlier or later the affected returns will show up in the reporting sequence. The method also rests on the fragile hope that the patched-up curve traced out by tonight's returns will flow parallel to the historical curve—an assumption that will not hold up if there is a differential shift in particular

⁹Jack Moshman, "Mathematical and Computational Considerations of the Election Night Projection Program," paper presented at the Spring Joint Computer Conference in Atlantic City, N.J., on May 2, 1968, p. 3.

areas to a particular candidate. For example, if areas that normally report late and also normally vote somewhat Democratic suddenly shift very strongly toward the Democratic candidate because of that candidate's special appeal in those areas, then the paths traced out by the historical curve and tonight's curve would not be parallel, and the projection might be wrong. Finally, the method does not easily accommodate new political factors, such as a third-party candidate.

Because of these limitations and the availability of more powerful, more inferentially secure methods, mu curves are not now widely used in electoral projections, although they do retain some utility for informal use in interpreting election returns. That utility comes from the limited insight upon which mu curves are based: that different areas, with different voting patterns, report their returns at different times on election evening. Of course we knew that anyway.

The second—and preferred—forecasting method compares tonight's returns from those counties (or wards, precincts, or the like) that have reported early with the returns from previous elections in those same counties. The adjustment of current returns by previous per-

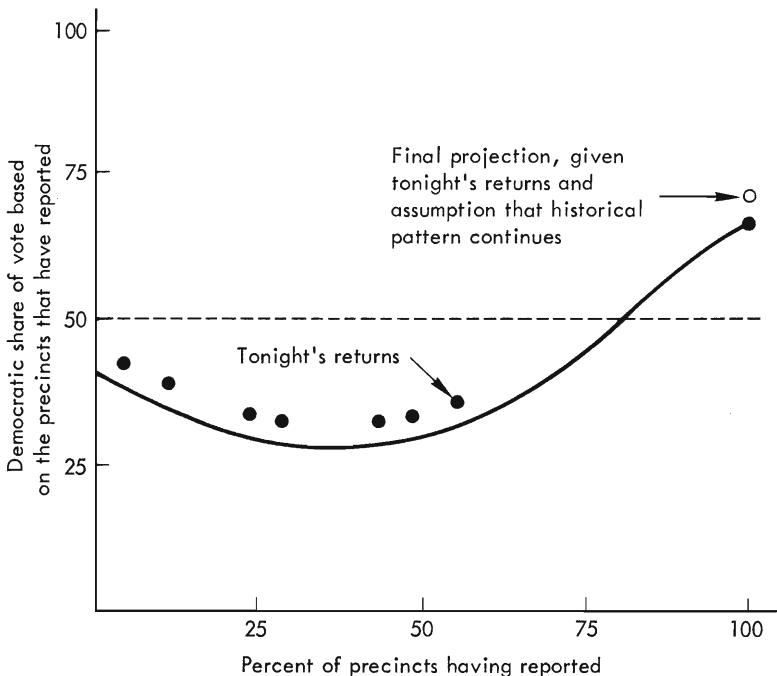


FIGURE 2-5 Comparing tonight's returns with the historical pattern to make a projection

formance at a disaggregated level (that is, at the county level) requires more detailed data and analysis than the mu-curve method—but it yields far more inferentially secure results. That is, there is a good chance that we know more after having done the analysis than we did before.

Comparing tonight's returns from a county with its previous voting patterns takes into account that the counties reporting first are not a representative sample. Counties with early complete returns may tend, in some states, to be Republican counties; in others, Democratic counties. At any rate, why hope they are typical or representative? Comparing current returns with old returns will adjust or control for a county's normal political leanings. For example, the raw returns from Massachusetts are not very helpful in projecting the national winner in a presidential race; but such returns are helpful if we know that Massachusetts normally runs heavily Democratic. So, if the Democratic candidate barely leads in Massachusetts, then that candidate is surely in real trouble nationwide.

Note the assumption here that the shift or the swing toward one party is roughly the same over the whole state or the whole nation. This assumption will not however lead to disaster—because it can be checked on election night with the data in hand simply by comparing the shifts across the counties that have reported. If the shifts are not consistent across counties, then either the historical base values from previous elections for the counties are ill-chosen and inappropriate for judging the pattern of tonight's election, or else the candidates had a special appeal to certain groups clustered by region and the shifts are not the same for different parts of the country. In contrast, violations of assumptions behind the mu-curve method are not easily discovered—at least in the short-run on election night.

Thus the second projection method is somewhat more powerful and safer than the use of mu curves because its assumptions are more modest and because some of its important assumptions can be verified during the course of the analysis. The second method does, however, require much more data and computing power; the grand assumptions of the mu curves are replaced by the collection and analysis of data.

In practice, the final projection of the election consists of a combination of several separate projections. This mixture forming the final, aggregate projection melds several component projections together:

1. the projection from the method of county-adjusted returns:
 $\%D_c$ = percent Democratic projected from counties;
2. the projection resulting from the so-called "key precincts," which are chosen either randomly or because of their special political

- interest: $\%D_k$ = percent Democratic projected from key precincts;
- the projection of the race before any returns are in at all, called a "prior"—a projection based on pre-election polls or political judgment: $\%D_p$ = prior projection of percent Democratic.

How much of each projection is mixed into the overall combined or "meld" projection? The prior, of course, receives full weight when no returns are in; as the returns pile up, the prior should carry less and less weight in the meld projection. Figure 2-6 shows one such weighting plan, with the weight, $w(r)$, a function of the number of precincts reporting. How should the other factors, $\%D_c$ and $\%D_k$, be weighted in the grand meld projection? Statisticians have a standard answer: form a weighted average using the reciprocal of the variances for weights.

Reciprocal weights are a reasonable choice—for, if the variance of an estimate is big, the weight should be small; if the variance of the estimate is small, then the estimate should have a relatively heavy weight and count for more because we have that estimate more precisely pinned down. Weighting by reciprocal variances gives, under ideal circumstances, the most precise combination. For the

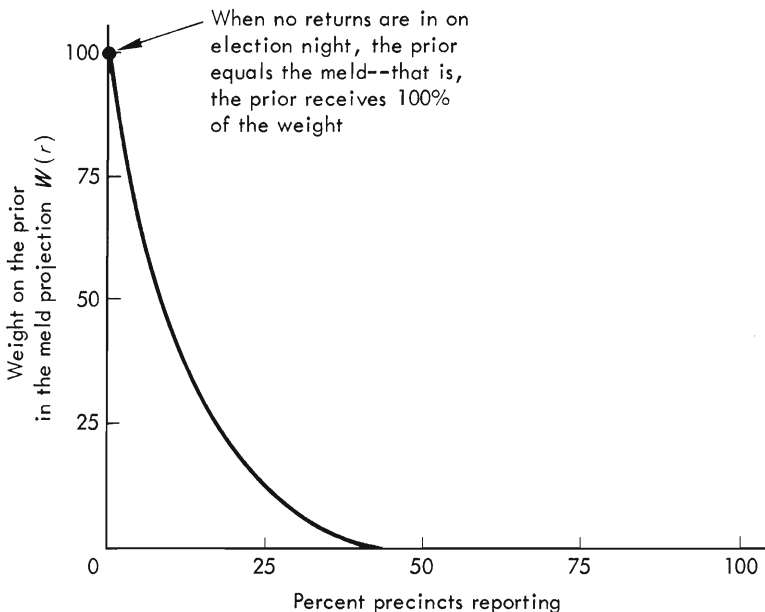


FIGURE 2-6 Weighting the prior in the overall meld

realities of election night, less simple combinations may be important. At any rate, one possible meld is the weighted average (weighted by the reciprocal of the variances) of the component projections:

$$\text{meld projection} = \frac{\frac{1}{S_c^2} \%D_c + \frac{1}{S_k^2} \%D_k + w(r) \%D_p}{\frac{1}{S_c^2} + \frac{1}{S_k^2} + w(r)}$$

where S_c^2 and S_k^2 are the variances of the estimates of $\%D_c$ and $\%D_k$. This is simply the particular realization of the general formula for a weighted average:

$$\text{weighted average} = \frac{\text{sum of weighted components}}{\text{sum of weights}} = \frac{\sum w_i x_i}{\sum w_i}.$$

Although based on the principles we have looked at here, contemporary projection models include many additional complications—complex estimation procedures, specially tailored base values, checks for bad data, and estimates of turnout. While today's elaborate models must be entirely computer based, in past years the votes were tabulated by hand on adding machines. Some years ago, the story has it, the truck delivering the dozens of rented adding machines to the studio on election day never arrived. Momentary panic arose, for how could they tabulate all the separate vote reports about to start pouring in? Finally, someone discovered a quickly available substitute for the adding machines. That night, ignoring the heavy-handed symbolism, they rang up the vote for president on cash registers!

Our next example evaluates another device for electoral forecasting—the “bellwether” district.

Bellwether Electoral Districts¹⁰

Time present and time past
Are both perhaps present in time future,
And time future contained in time past.

—T. S. Eliot, *Four Quartets**

¹⁰This section was co-authored with Richard A. Sun.

*From *Four Quartets* by T. S. Eliot. Reprinted by permission of the publishers, Harcourt Brace Jovanovich, Inc. and Faber and Faber Ltd.

Prior to the 1936 presidential election, the conventional political wisdom had it that as Maine voted, so went the rest of the nation. After the 46-state landslide, James Farley, Roosevelt's campaign manager, revised the theory: "As goes Maine, so goes Vermont." Such is perhaps the inevitable fate of so-called bellwether or barometric electoral districts; still, there are always new contenders with markedly unblemished records of retrospective accuracy to replace wayward bellwethers. Given the familiar inferential caution that retrospective accuracy provides little guarantee of prospective accuracy, what is the worth of claims that certain districts invariably reflect the national division of the vote?

The answers at hand differ: a skeptical statistician probably has little faith in the after-the-fact predictive success of bellwether districts; the collector of political folklore marvels at the record of such byways as Palo Alto County (Iowa) and Crook County, (Oregon) which have voted for the winner of every presidential election in this century; the newspaper reporter interviews a few citizens of Palo Alto or Crook County in search of "clues as to what will happen next Tuesday"; and Louis Bean has written four books premised on the notion that as goes *X*, so goes the country.¹¹ Here we will examine the question more deeply—and, at the same time, see a number of fundamental statistical techniques in action.

The data for the analysis are the election returns from almost all 3100 U.S. counties for the fourteen presidential elections from 1916 to 1968.¹² We will be looking for what are called "all-or-nothing" bellwethers: the county either votes for the winner of the presidential election or it does not. This seems to be the usual meaning of "bellwether district"; most discussions of supposed bellwethers report that the district has voted with the winner in the last *N* elections. Sometimes

¹¹ *Ballot Behavior* (Washington, D.C.: Public Affairs Press, 1940); *How to Predict Elections* (New York; Knopf, 1948) *How America Votes in Presidential Elections* (Metuchen, N.J.: Scarecrow Press, 1968); and *How to Predict the 1972 Election* (New York: Quadrangle, 1972).

¹² The data tapes were made available through the Inter-University Consortium for Political Research. We edited them extensively, correcting errors and adding missing data. Of the 3070 counties in the United States, we have the complete two-party election returns for the fourteen elections from 1916 to 1968 for 2938 counties, or 96 percent. The remaining counties had to be dropped because one election in the fourteen election series was missing; others may have changed names or are mixed in with other political units. A listing of the missing counties and election years was reviewed both before and after our analysis; both times it appears that the small amount of missing data had no consequences for our findings. Some of our early computations carried along votes for four different parties in each county, but we finally edited the data to include only the returns for the two major parties. Therefore all election returns reported here are based on the votes of the two major parties in all the elections.

N is surprisingly small; some journalists have interviewed nonrandomly selected citizens of "bellwether" communities that have voted for the winner in only three or four previous elections.

One good test of the credibility of bellwethers is to conduct a series of historical experiments, each designed to answer the question: How well would we have done in predicting the election of 19XX if we had followed a group of supposedly bellwether counties chosen on the basis of past elections before the election of 19XX? For example, going into the 1968 election, there were 49 counties that had voted for the winner in every presidential election since 1916—thirteen elections (or more) in a row with the winner. Were these 49 retrospective bellwethers more likely than other counties to support the winner in 1968? This is the sort of question that we will answer over and over, for different elections and for different choices of historical bellwethers.

Since they directly answer the question at hand, the historical experiments seem to provide the most powerful means of assessing the credibility of bellwethers. It is also possible to construct probability models to provide a baseline or null hypothesis against which to compare the observed performance of reputed bellwethers. We met with little success in developing models based on reasonable assumptions. The construction of a useful probability model remains an open question, although we suspect that even a very good model would still not provide as direct and powerful test of bellwethers as the historical experiment.

Another statistical problem arises because bellwethers are found in an after-the-fact search through election returns; there is no theory identifying particular areas as potential bellwethers before the fact. We have then a situation analogous to that of "shotgunning" in survey research: the searching through of a large body of data for statistically significant results leads to difficulties in just how to include the fact of the search in an adjusted significance test. One answer is simply the independent replication on a fresh collection of data of the results found through searching. That is, of course, the underlying logic of the historical experiment: bellwethers are chosen from a search, and then we see if their bellwether performance is replicated in the historical future.

The usual technique for evaluating bellwethers is retrospective admiration of the historical record. Almost all written accounts of reputed bellwethers describe an area's lengthy record in voting for winners and then ask, in effect, "Isn't that something?" These accounts evaluate the predictive performance of the past without reference to either prospective accuracy or the predictive record of other areas. Consider excerpts from a typical *New York Times* story on bellwethers:

Town Votes 'Em As It Sees 'Em
And It Usually Sees 'Em Right

Salem, N.J., April 8—The political professionals are keeping an eye on this small Quaker community in southern New Jersey for clues to the outcome of the presidential election.

For fifty years, with only two exceptions, Salem has voted for the victorious Presidential candidate. . . .

There is no clear reason for Salem's stature as an election indicator.

"But," says County Clerk Thomas J. Grieves, "you can't call it chance or a quirk. It happens too often. . . ." ¹³

Actually, there are several hundred counties with predictive records better than Salem's over the last fifty years. But the important point is that no evaluation of Salem's record can be made on the basis of past election returns from Salem alone. A bellwether's credibility can only be assessed by examining, in comparison to other districts, its *predictive* record and not merely its *postdictive* record.

Consider the following historical experiment: let us choose the counties with the best records for predicting presidential elections from 1916 to 1964 and see how well they predicted the outcome of the 1968 election. There were 49 such counties with records of supporting the winner in all 13 elections from 1916 to 1964. Such a record, by almost any standard, is a bellwether performance—if the counties had been identified in 1916 instead of after the fact. How well did the 49 retrospective bellwethers of 1916–1964 do in predicting the winner in 1968? Not very well at all; 27 of the 49 (or 55.1 percent) voted with the winner in 1968. Two-thirds of *all* counties supported the winner in 1968, and so a county chosen at random could typically have been expected to outpredict the counties with previously perfect predictive records. Table 2-3 shows the full array of results, with the 1968 predictive performance tabulated against the prior record of predictive accuracy. Oddly enough, the best predictions in 1968 were made by counties that had had the worst record in the past (5 right, 8 wrong). These 80 counties (that went 100 percent for the winner in 1968) were, of course, counties that had voted without fail for the Republican candidate in every previous election since 1916 and persisted in 1968. So it is easy to find a group of counties, identified by their past voting record, that will support the upcoming winner—if you only know how the election is going to turn out!

The election of 1968 was a particularly bad year for the bellwethers of the past. Table 2-4, repeating the tests for the presidential elections from 1936 to 1964, shows that for some elections the bellwethers

¹³ *The New York Times*, April 9, 1964, p. 29.

TABLE 2-3

Predictive Performance from 1916 to 1964 Compared with Predictive Record in 1968 Election

Past Performance, 1916-1964				1968 Performance			
Past Predictions		Counties		Right		Wrong	
Right	Wrong	Number	Percent	Number	Percent	Number	Percent
0	13	0	0.0	0	0.0	0	0.0
1	12	0	0.0	0	0.0	0	0.0
2	11	0	0.0	0	0.0	0	0.0
3	10	0	0.0	0	0.0	0	0.0
4	9	0	0.0	0	0.0	0	0.0
5	8	80	2.7	80	100.0	0	0.0
6	7	229	7.8	209	91.3	20	8.7
7	6	502	17.1	303	60.3	199	39.6
8	5	708	24.1	424	59.9	284	40.1
9	4	554	18.8	397	71.6	157	28.3
10	3	380	12.9	251	66.0	129	33.9
11	2	274	9.3	148	54.0	126	45.9
12	1	162	5.5	97	59.9	65	40.1
13	0	49	1.6	27	55.1	22	44.9
		2938	100.0	1936	65.9	1002	34.1

of the past do predict the upcoming election somewhat more accurately than a typical county.

Tables 2-3 and 2-4 provide us with a great deal of experience with retrospective all-or-nothing bellwethers. The tables suggest:

1. Perhaps each time one hears of an area with a spectacular predictive record in the past, a glimmer of hope and curiosity arises suggesting that surely this fine record couldn't be mere chance—there must be *something* going on. Whatever that something might be, it isn't a high degree of prospective accuracy. Sometimes previously accurate districts do better than just any collection of districts; sometimes they don't. The retrospective bellwethers were particularly poor in the close elections of 1960 and 1968. The compilations of Table 2-4 show the erratic record of the retrospective all-or-nothing bellwethers in predicting the future.

2. We have identified "bellwethers" in Tables 2-3 and 2-4 by their previously perfect predictive records in at least six consecutive previous elections. If this standard is applied to judging the results of our historical experiment, then the bellwethers of the past are not the bellwethers of the present. In five of the eight elections, the previously bellwether counties had a higher probability of voting with the winner

TABLE 2-4
 Predictive Record of Previously Accurate Counties in Presidential Elections,
 1940-1964

PREDICTING 1940	<i>Number of counties</i>	<i>Percent voting with winner, 1940</i>
1916-1936 past performance, right-wrong = 6-0	602	52.9
Nationwide	2938	61.6
PREDICTING 1944	<i>Number of counties</i>	<i>Percent voting with winner, 1944</i>
1916-1940 past performance, right-wrong = 7-0	319	72.7
Nationwide	2938	55.3
PREDICTING 1948	<i>Number of counties</i>	<i>Percent voting with winner, 1948</i>
1916-1944 past performance, right-wrong = 8-0	232	87.5
Nationwide	2938	59.9
PREDICTING 1952	<i>Number of counties</i>	<i>Percent voting with winner, 1952</i>
1916-1948 past performance, right-wrong = 9-0	203	81.3
Nationwide	2938	68.3
PREDICTING 1956	<i>Number of counties</i>	<i>Percent voting with winner, 1956</i>
1916-1952 past performance, right-wrong = 10-0	165	87.3
Nationwide	2938	70.0
PREDICTING 1960	<i>Number of counties</i>	<i>Percent voting with winner, 1960</i>
1916-1956 past performance, right-wrong = 11-0	144	35.4
Nationwide	2938	38.6
PREDICTING 1964	<i>Number of counties</i>	<i>Percent voting with winner, 1964</i>
1916-1960 past performance, right-wrong = 12-0	51	96.1
Nationwide	2938	73.3

than a county chosen at random from the nation as a whole; in the other three elections (1940, 1960, and 1968), a county chosen at random would be the county of choice in predicting the upcoming election.

3. The retrospective bellwethers, *taken as a group*, correctly predicted seven of the eight trial elections—in the sense that a majority of the group of retrospective bellwethers supported the winner. Exactly the same was true of a group of randomly selected counties (within the limits of sampling error).

4. There were, alas, no anti-bellwether counties. No county had such an outstandingly poor record that it could serve, by reversing its preferences, as a predictive (or even postdictive) guide.

5. Tables 2-3 and 2-4 indicate clearly why one obvious probability model, the binomial, for all-or-nothing bellwethers does not provide a useful baseline. Consider the following: if a fair coin, labeled "Democratic candidate will win" on one side and "Republican candidate will win" on the other, were tossed prior to each of the last 14 presidential elections, the probability that the coin would successfully predict the winner of all 14 contests is

$$\left(\frac{1}{2}\right)^{14} = \frac{1}{16,384} = .000061.$$

If this toss of the coin were performed in each of the 3100 counties, then it would be expected that

$$(.000061)(3100) = 0.2 \text{ counties}$$

would correctly go along with the winner 14 elections in a row. More generally, the binomial model for k successes in 14 independent trials with probability of success equal to one-half generates the distribution of predictions shown in Figure 2-7. The actual distribution of counties is also shown in the figure. It is clear that the distribution of actual election outcomes is not generated by a process of 14 independent trials with probability of success equal to one-half. That is because the probability of success usually substantially exceeds one-half and the trials are, in fact, highly dependent. The chances that a given county votes with the winner is usually around two-thirds, as Tables 2-3 and 2-4 show.

A more difficult problem in constructing a probability model is that the election results are not independent over space and time: both the interelection and intercounty correlations are very high. For example, the correlation between the division of the vote from one election to the next over all counties is almost always greater

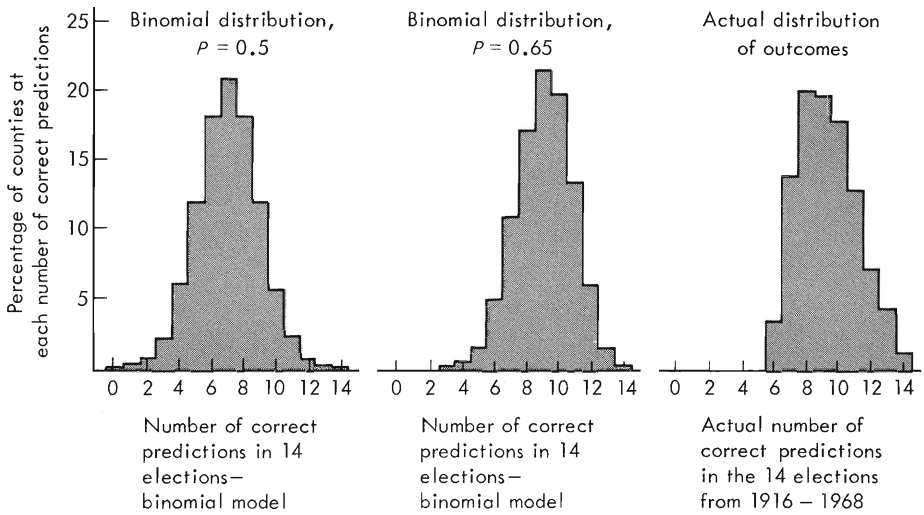


FIGURE 2-7 Binomial and actual outcome distributions

than .90. Considering that a county could go either Democratic or Republican in each of the 14 elections yields $2^{14} = 16,384$ theoretically possible electoral histories or paths that the counties could have followed over the 56 years. Less than 400 of these electoral histories actually occur, and only about 30 contain more than a handful of counties. At least 40 percent of all counties have gone more or less straight Democratic or straight Republican with occasional deviations in landslide years (Table 2-5).

TABLE 2-5
Most Frequently Occurring County Electoral Histories, 1916-1968

<i>History</i>	<i>Number of counties</i>
Straight Democratic	200
Democratic, except 1964	160
Democratic, except 1968	54
Democratic, except 1964 and 1968	58
Straight Republican	79
Republican, except 1964	128
Republican, except 1932, 1936, and 1964	136
Republican, except 1916, 1932, 1936, and 1964	155
Followed nation, all elections	27
Followed nation, except 1960	68

6. Twenty-seven of the nation's 3100 counties voted for the winner in every presidential election from 1916 to 1968. It may be possible—or at least a firm believer in bellwethers might well argue—that there are some truly bellwether districts hidden in those counties. What we have shown, of course, is only that counties with perfect postdictive records have undistinguished predictive records—when those counties are *taken as a group*. The only way we can identify bellwethers is as members of such a group. One final shred of evidence is to consider the performance of the nation's finest bellwethers. Prior to the 1960 election, there were eight counties in the nation with records of supporting every winner in this century. After 1968, only three of these eight superbellwethers still had unblemished records: Crook County, Oregon; Laramie County, Wyoming; and Palo Alto County, Iowa. They remained accurate in 1972.

Our conclusion in the case of all-or-nothing bellwethers is clear: the usual concept of a bellwether electoral district has no useful predictive properties. The all-or-nothing counties are only a curiosity and probably should be forgotten. It is a waste of time to send reporters out to interview nonrandomly selected citizens of Crook County a week or two before the election—at least it is a waste of time from any sort of scientific point of view. Such news reports create mystery where little exists.

There perhaps remains a magical air about the bellwethers of the past; some of these districts, considered individually, seemingly have such phenomenal records and yet we know better than to take them seriously—but still. . . . It may be best to look not to the election returns for the source of the mystery, but rather to ourselves. Maugham once wrote:

The faculty for myth is innate in the human race. It seizes with avidity upon any incidents, surprising or mysterious, in the career of those who have distinguished themselves from their fellows, and invents a legend to which it then attaches a fanatical belief. It is the protest of romance against the commonplace of life.¹⁴

¹⁴Somerset Maugham, *The Moon and Sixpence* (Harmondsworth, Middlesex, England: Penguin Books, 1941), p. 7.

Regression Toward the Mean: How Prior Selection Affects the Measurement of Future Performance

Consider the defects in research design in the following example:

Students in a statistics course who needed remedial teaching (as indicated by their performance in the lower quartile of an achievement test in arithmetic) were assigned to a special class in sensitivity training. Soon the teacher of the special class was able to go into full-time educational consulting because of the success of his new book, *Ending Educational Hangups in Statistics: How Empathy Pays Off*. The book showed that the special class was strikingly effective because when the students in the special class took the tests again after only six months, their test scores had greatly increased—in fact, almost all the way up to the average of the first test scores of all the students who initially took the arithmetic test.

Several difficulties that are common in research designs compromise this hypothetical example.

This design uses the first test to divide the class into a treatment group (consisting of the lower quartile of students) and a control group (the remainder of the class). Students in the treatment group took the same tests again six months after joining the special class. The following comparisons were made in an effort to assess the benefits of the special class:

1. Average “gain” for special class equals

$$\left(\begin{array}{l} \text{average of scores on} \\ \text{second test for special} \\ \text{class} \end{array} \right) \text{ minus } \left(\begin{array}{l} \text{average of scores on} \\ \text{first test for special} \\ \text{class} \end{array} \right)$$

2. “Improvement” relative to rest of class equals

$$\left(\begin{array}{l} \text{average of scores on} \\ \text{second test for special} \\ \text{group} \end{array} \right) \text{ minus } \left(\begin{array}{l} \text{average of scores for} \\ \text{whole class on the first} \\ \text{test} \end{array} \right)$$

Two serious defects in the research design result in a bias in the "gain" and "improvement" scores such that the beneficial effect of the special class is exaggerated. The first defect is the failure to take into account the effect of practice and maturation on the test scores. Students taking a test a second time, as in the special class, can be generally expected to get better at taking tests; consequently, their scores improve merely because of their increased experience. Similarly, since the treatment-group scores on the second test are compared with the earlier test scores of the control group, a bias due to the maturation of the special group results. In other words, the students in the special group may improve relative to their previous performance (and the previous performance of their contemporaries) merely because they are older and smarter and not because they are necessarily benefiting from the special class.

In this design, then, the improvements in the scores of the special group due to practice and maturation effects are incorrectly attributed to the effect of the special class. Although it is impossible without additional information (or a better research design—see below) to judge the exact strength of the bias, we do at least know its direction: it favors the hypothesis that there is benefit from the special class.

The second defect in the research design is more subtle. It is a version of what is called the "regression fallacy." If members of a group are selected because their scores are extreme (either high or low) on a variable and if this extreme group are later tested once again, we will generally find that the group are "more average" than they were on the first test. Their scores will have moved or "regressed" toward the mean. One way to view the situation is to think of the extreme group as consisting of two sorts of people: (a) those who deserve really to be in that group and (b) those who are there because of random error—unlucky guesses on the test, an "off" day, and so forth. When the extreme group is tested a second time, the group (b) will typically perform more like their true selves, thereby raising their scores on the average at least. The deserving extremists in group (a) will continue their poor scores, albeit with some variation.

Thus the average score of the extreme group will typically increase because of the more typical performance of group (b) on the second test. There is no way of distinguishing group (a) from group (b) with only one test.

The problem arises when any group is formed by selecting its members because they are extreme on a single measure. For example, let us say that the highest quartile of students were placed in the special class instead of the bottom quartile. What would happen then? Once again, two types of students make up the extreme top group:

(a) those who are actually skilled and who deserve to be placed in the top quartile and (b) those who are lucky, who guess right, and so on. Now if this group is tested once again, it will generally be found that the overall average of the original extreme group has dropped somewhat—because not all the lucky performers on the first test will be lucky again.

The fallacy occurs in all sorts of situations. Wallis and Roberts provide several good examples including the following:

Teachers—except, of course, statistics teachers—sometimes commit the regression fallacy in comparing grades on a final examination with those on a midterm examination. They find that their competent teaching has succeeded, on the average, in improving the performance of those who had seemed at midterm to be in precarious condition. This accomplishment naturally brings the teacher keen satisfaction, which is only partially dampened by the fact that the best students at midterm have done somewhat less on the final—an “obvious” indication of slackening off by these students due to overconfidence.¹⁵

Let us examine a numerical example of what might have happened in the case of the special class. Make the following assumptions:

1. There are no practice or maturation effects.
2. The special class has no effect at all on the students' test scores.

Under these assumptions we should observe no significant gains or improvements by the special class if the research design is free of bias. If, however, the research design has a bias, we will be able to get at least an approximate idea of its extent. Table 2-6 shows three sets of made-up test scores:

Column I: *The “true score” of each student on the test.* This, of course, is never actually measured perfectly, and the remaining columns represent the true score plus some random measurement error.

Column II: *The “true score” for each student with a random number between -20 and 20 added to each score.*

Column III: *Again the “true score” with another random number added to column I.*

Let the numbers in column II represent the scores of all the students on the first test and those in column III the scores on the second test. Since the test scores were computed by adding a random error to the “true scores,” we find that there is very little difference in

¹⁵W. Allen Wallis and Harry V. Roberts, *Statistics: A New Approach* (New York: Free Press, 1956), p. 262.

TABLE 2-6
Random Errors Added to True Scores

<i>Student</i>	<i>I</i>	<i>Random error, test 1</i>	<i>II</i>	<i>Random error, test 2</i>	<i>III</i>
	<i>True score</i>		<i>Observed score, test 1</i>		<i>Observed score, test 2</i>
A	70	+13	83*	+1	71
B	75	-20	55*	+15	90
C	80	+8	88	-13	67
D	84	+7	91	-1	83
E	87	-15	72*	-9	78
F	90	+2	92	+8	98
G	93	-4	89	+12	105
H	95	-7	88	+16	111
I	96	+3	99	-12	84
J	97	+17	114	+20	117
K	98	-19	79*	-1	97
L	99	+11	110	+5	104
M	99	-18	81*	-17	82
N	100	-13	87*	+3	103
O	100	+9	109	-7	93
P	101	+12	113	+10	111
Q	101	-0	101	-5	96
R	102	-18	84*	+2	104
S	103	+13	116	+9	112
T	104	+7	111	-15	89
U	105	+3	108	+14	119
V	107	+12	119	-7	100
W	110	-11	99	+16	126
X	113	-20	93	+5	118
Y	116	+15	131	-19	97
Z	120	+1	121	+5	125
AA	125	-2	123	-2	123
BB	130	-14	116	-14	116

*The asterisk indicates students in lowest quartile on test 1.

the average score of the whole class on test 1 compared with test 2. Also the test seems to be measuring something: the correlation between the tests is .51. The correlation would be perfect, if we had not introduced the random measurement error into the true score on each test. Furthermore, note that the variability on both tests 1 and 2 is the same.

It should be clear that all that has been done is to construct some test scores containing some random error. No systematic effects in the data enable one to differentiate between the results of test 1 and test 2. But let us now see what happens in the research design

used in assessing the effects of the special class. The students in the special class were chosen because they were in the bottom of the class on the first test. Compare, then, the scores of the lowest seven students in the class as measured by test 1 (Table 2-7).

This research design generates the following misleading results. The average score of the group entering the special class was 77.3; after attending the special class for six months, their average score was 89.3—a “gain” of 12.0 points. Thus, because of the regression effects operating in this research design, a *pseudo-gain of 12 points* was found between test 1 and test 2, even though all the difference between test 1 and test 2 was generated by random numbers.

Note how plausible it all seems. A group of students are selected on the basis of test scores to enter the special class, and when the same students are tested later, those in the special class appear to have gained 12 points. Test 1 and test 2 are rather highly correlated, indicating that the tests are moderately reliable. And yet it is all a statistical artifact.

What would be a better research design—one that assesses the effect, if any, of the special class but avoids the bias resulting from the effects of practice, maturation, and regression toward the mean? The essential feature of an improved research design is that not all of the low scorers should be placed in the special group. Ideally, some of the low scorers on test 1 should be randomly assigned to the special group; the others should remain in the regular class. In evaluating the effects of the special class, then, the basic comparison should be made between those low scorers in special class versus those low scorers in the regular class. Regression toward the mean still operates in this design, but its impact is roughly equal on the

TABLE 2-7

Scores on Test 1 Compared to Scores on Test 2 for the Lowest Quartile of Students on Test 1: Pseudo-Gains and Pseudo-Losses

<i>Student</i>	<i>Test 1</i>	<i>Test 2</i>	<i>Difference: "Gain" > 0 "Loss" < 0</i>
A	83	71	-12
B	55	90	35
E	72	78	6
K	79	97	18
M	81	82	1
N	87	103	13
R	84	104	20

control group and the treatment group because students were randomly assigned to the two groups.

The improved design, however, does give us a chance to separate out the genuine effects resulting from membership in the special class from the artifactual effects deriving from practice, maturation, and regression toward the mean. The original design confounds these factors and throws them all into the gain score.

This example also illustrates the utility of trying out the design and analysis on realistic but random data. Random data contain no substantive effects; thus if the analysis of the random data results in some sort of effect, then we know that the analysis is producing that spurious effect, and we must be on the lookout for such artifacts when the genuine data are analyzed.

Prediction of Accident Proneness: Can Producers of Automobile Accidents Be Identified in Advance as Consumers of Traffic Violations?

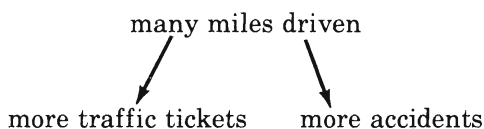
Only a small number of drivers are involved in severe automobile accidents. This fact gives rise to statements like "Three percent of all drivers produce one hundred percent of all severe accidents." The statement, while true, can be misleading. It does not mean that a small group of drivers go around systematically running down people or ramming other cars. "Accident proneness" may or may not be a useful concept.

It is empirically true that a small number of people, not necessarily identifiable in advance, are involved in serious accidents. Do these people have any characteristics in common? Can we ascertain roughly the probability that a given driver will be involved in an accident within a certain period of time? Insurance companies already make such predictions in a crude way by setting their rates in relation to factors including the driver's age, sex, marital status, accident history, type of driving, and record of traffic violations. Such procedures, at least as they are employed in Canada, are biased against some drivers (particularly high-risk drivers) because the various factors are not independent, resulting in double counting of risks against some drivers.¹⁶

¹⁶See R. A. Holmes, "Discriminatory Bias in Rates Charged by the Canadian Automobile Insurance Industry," *Journal of the American Statistical Association*, 65 (March 1970), 108-22.

A study of the relationship between the number of traffic violations a driver collects and his or her involvement in accidents is threatened by possible spurious correlations. First, one result of a motor vehicle accident is a traffic ticket. One driver or another is found to have committed a violation which “explains” the accident. This leads to statements such as “Accidents are caused by excessive speed,” which are based on evidence that in many accidents, drivers involved are adjudged to have exceeded the speed limit. Lacking here is a comparison group of the speed of drivers *not* involved in accidents. There is some evidence that a large proportion of all drivers on the road are, in fact, exceeding the speed limit. In any case, a first step in a study of traffic violations and accidents is to control for the tickets produced by accidents—at least if the task is to predict, on the basis of a past history of traffic violations, that certain drivers will be more likely to be involved in accidents.

A second problem of potential spuriousness is suggested by the following model:



Thus, high-mileage drivers face greater exposure to the risk of both a traffic ticket and an accident—even if they drive with a care equal to that of low-mileage drivers.

A review of the studies of the relationship between violations and accident involvements points to both of these problems and to a partial solution:

Ross investigated the relationship between violations and accidents for the 36 accident-involved drivers . . . and found that 12 of these 36 drivers had reported traffic convictions on their official records. These 12 people had 18 convictions. However, since there was no control group in this study, it is not possible to ascertain whether drivers with accidents had a higher violation rate than drivers without accidents. A point made by Ross, and one which has an important bearing on other studies using official records or information collected in interviews, is that there were discrepancies between interviewee-reported and recorded accidents and violations large enough to throw question upon studies relying on one or the other source of information in arriving at an accident or violation record.

As part of a California driver record study, relationships between concurrent recorded accidents and citations (convictions for moving traffic violations) were analyzed. The data for this analysis consisted of a random sample of 225,000 out of approximately eleven million existing California driving records. Each driving record included a three-year history of both accidents and citations. To avoid inadvertent

correlation effects, citations directly resulting from accident investigations were labeled as "spurious" and were removed from the citation counts in most of the analysis.

The driver records were grouped according to the number of nonspurious citations, and the mean number of accidents per 100 drivers was calculated for each group. This analysis indicated an approximately linear relationship between citations and accidents with fluctuations at the high end of the citation count scale as a result of reduced sample size. Whereas those with no countable citations in the three-year period had only 14 accidents per 100 individuals, those with five citations had 62 accidents per 100 individuals and those with nine or more citations had 89 accidents per 100 individuals.

These figures indicate that there is a strong relationship between the mean number of accidents per driver and the number of concurrent citations when large groups of drivers are considered. On the other hand, the correlation coefficient between accidents and nonspurious citations was only 0.23. This low figure indicates that large errors could be made if one attempted to estimate the number of accidents an individual driver had on the basis of his citation record over the same time period. One would generally expect the correlation between concurrent events to be higher than nonconcurrent events. Thus, one should expect even larger errors, if one attempted to predict an individual's future accident record on the basis of his past citation record.

High-mileage drivers, other factors being equal, are exposed to a higher risk of both accidents and citations. Variations from driver to driver in exposure in general and annual mileage in particular may produce part of the correlation between accidents and citations that has been observed. Another California study examined characteristics of negligent drivers, defined as those whose record indicated a point count of four or more in 12 months, of six or more in 24 months, or eight or more in 36 months. (A point is scored for each traffic violation involving the unsafe operation of a motor vehicle or accident for which the operator is deemed responsible; two points are scored for a few types of violations deemed especially serious.)

When the annual mileage for a group of negligent drivers over age 20 was compared with that for a random sample of renewal applicants it was found that the negligent group averaged 17,219 miles per year while the applicant group averaged 7,449 miles per year. When males and females were treated separately it was found that negligent males averaged 17,591 miles per year as contrasted to 9,649 miles per year for the male applicants, while negligent females averaged 9,403 miles per year as contrasted to 5,519 miles per year for female applicants. The negligent drivers may have inflated their reported annual mileage in order to impress officials with their need to drive; nevertheless, it appears very likely that the negligent drivers do indeed drive more than average.¹⁷

¹⁷ *The State of the Art of Traffic Safety*, by Arthur D. Little, Inc., for the Automobile Manufacturers Association, Inc. (Cambridge, Mass.: Arthur D. Little, Inc., June 1966) pp. 42-43.

Spellbinding Extrapolation

One of the most spellbinding efforts at simple extrapolation beyond the data arises in this history of guano:

Guano, as most people understand, is imported from the [islands of the] Pacific—mostly of the Chincha group, off the coast of Peru, and under the dominion of that government.

Its sale is made a monopoly, and the avails, to a great extent, go to pay the British holders of Peruvian Government bonds, giving them, to all intents and purposes, a lien upon the profits of a treasure intrinsically more valuable than the gold mines of California. There are deposits of this unsurpassed fertilizer, in some places, to the depth of sixty or seventy feet, and over large extents of surface. The guano fields are generally conceded to be the excrements of aquatic fowls, which live and nestle in great numbers around the islands. They seem designed by nature to rescue, at least in part, that untold amount of fertilizing material which every river and brooklet is rolling into the sea. The wash of alluvial soils, the floating refuse of the field and forest, and, above all, the wasted materials of great cities, are constantly being carried by the tidal currents out to sea. These, to a certain extent at least, go to nourish, directly or indirectly, submarine vegetable and animal life, which in turn goes to feed the birds, whose excrements in our day are brought away by the ship-load from the Chincha Islands.

The bird is a beautifully arranged chemical laboratory, fitted up to perform a single operation, viz.: to take the fish as food, burn out the carbon by means of its respiratory functions, and deposit the remainder in the shape of an incomparable fertilizer. But how many ages have these depositions of seventy feet in thickness been accumulating!

There are at the present day countless numbers of the birds resting upon the islands at night; but, according to Baron Humboldt, the excrements of the birds for the space of three centuries would not form a stratum over one-third of an inch in thickness. By an easy mathematical calculation, it will be seen, that at this rate of deposition, it would take seven thousand five hundred and sixty centuries, or seven hundred and fifty-six thousand years, to form the deepest guano bed. Such a calculation carries us back well on towards a former geological period, and proves one, and perhaps both, of two things—first, that in past ages, an infinitely greater number of these birds hovered over the islands; and secondly, that the material world existed at a period long anterior to its fitness as the abode of man. The length of man's existence is infinitesimal, compared with such a cycle of years; and the facts recorded on every leaf of the material universe ought, if it does not, to teach us humility. That a little

bird, whose individual existence is as nothing, should, in its united action, produce the means of bringing back to an active fertility whole provinces of waste and barren lands, is one of a thousand facts to show how comparatively insignificant agencies in the economy of nature produce momentous results.¹⁸

Rather substantial inferences, given the observed data!

¹⁸*London Farmer's Magazine: Prospectus of the American Guano Company* (New York: John F. Trow, 1855).

Two-Variable Linear Regression

“Yet to calculate is not in itself to analyze.”

—Edgar Allen Poe, *The Murders in the Rue Morgue*

Introduction

Fitting lines to relationships between variables is the major tool of data analysis. Fitted lines often effectively summarize the data and, by doing so, help communicate the analytic results to others. Estimating a fitted line is also the first step in squeezing further information from the data. Since the observed value can be broken up into two pieces,

observation = fitted value + residual,

we can therefore find the remaining part of the observed value that is unexplained,

residual = observation – fitted value,

and work with the residuals to discover a more complete explanation of the influences on the response variable.¹ Such was the procedure used in the study of automobile safety inspections in Chapter 1.

¹This follows J. W. Tukey and M. B. Wilk, “Data Analysis and Statistics: Techniques and Approaches,” in E. R. Tufte, ed., *The Quantitative Analysis of Social Problems* (Reading, Mass.: Addison-Wesley, 1970), pp. 373–74.

We now briefly review the mechanics of linear regression. The equation of a straight line is

$$Y = \beta_0 + \beta_1 X,$$

where β_0 is the intercept and β_1 is the slope as shown in Figure 3-1. The observed data are used to estimate the two parameters, β_0 and β_1 , of the model. The actual numerical *estimates* of the intercept and the slope are written as $\hat{\beta}_0$ and $\hat{\beta}_1$, where the “hats” indicate that the quantity is an estimate of a model parameter—an estimate that is computed from the observed data.

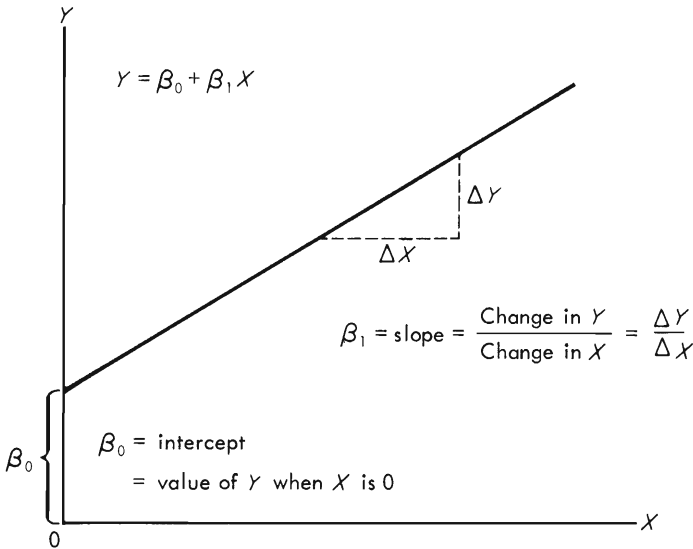


FIGURE 3-1 Equation of a straight line

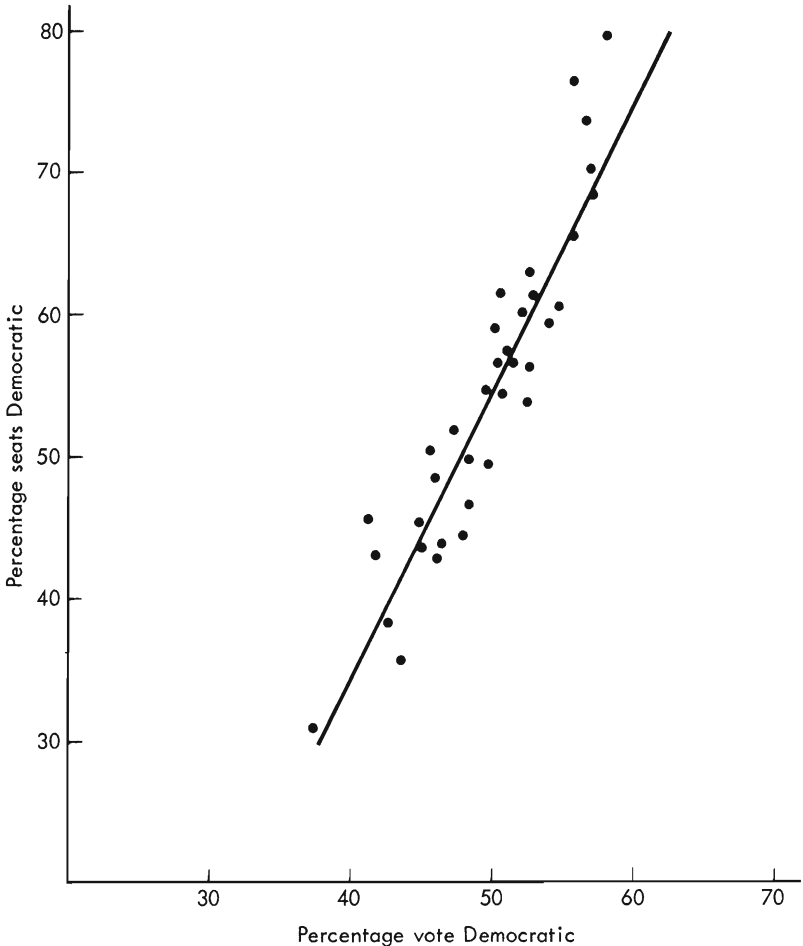
The slope, a summary of the relationship between X and Y , answers the question: when X changes by one unit, by how many units does Y change? The answer is that Y changes by β_1 units. Consider the following example. In the 36 congressional elections from 1900 to 1972, the line (shown in Figure 3-2)

$$\% \text{ seats Democratic} = -49.64 + 2.07 (\% \text{ votes Democratic})$$

fits the relationship between the share of congressional seats won by the Democrats and the share of votes that party received nationwide

for their congressional candidates. The estimated slope, $\hat{\beta}_1$, is 2.07; that is,

$$\hat{\beta}_1 = \frac{\text{change in } Y}{\text{change in } X} = \frac{\text{change in percent of seats}}{\text{change in percent of votes}} = 2.07.$$



Percentage seats Democratic = $-49.64 + 2.07$ (Percentage votes Democratic)

$$Y = -49.64 + 2.07X$$

$N = 37$ Congressional elections, 1900-1972

FIGURE 3-2 Fitted line and observed data

This means that a one percent change in the share of the Democratic vote was typically accompanied by a change of 2.07 percent in the Democratic share of seats in Congress. Thus an increase of only one percent in the share of the vote was worth a substantially larger increase (of a little over two percent) in the share of seats. Of course, it works the other way, too: a drop of one percent of the vote is associated with a loss of two percent of seats. Figure 3-2 shows the data and the fitted line. In this *particular* case, the estimate of the slope measures what is called the “swing ratio”—the swing or change in seats for a given change in votes. Often, then, the substance of the problem gives a special meaning to the slope, even though the mechanics of computing the slope are the same in each case.

The estimates of the slope and the intercept are chosen so as to minimize the sum of the squares of the residuals from the fitted line. This is the principle of *least squares*, which says

$$\text{minimize } \Sigma e_i^2,$$

$$\text{—that is, minimize } \Sigma (Y_i - \hat{Y}_i)^2$$

in the notation of Figure 3-3.

One of the glories of the principle of least squares is that it leads immediately to specific instructions as to how to use the data to compute $\hat{\beta}_0$ and $\hat{\beta}_1$ such that they uniquely satisfy the principle. The mathematics are found in any statistics text, where it is proved that the least-squares estimates of the slope and the intercept are computed from the observed data by

$$\hat{\beta}_1 = \frac{\Sigma (X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma (X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$

The fitted line minimizes errors in prediction when *X* is used to predict *Y*—and the errors in prediction are measured with respect to the *Y* variable. The estimate of the slope in this case is the *slope of the regression of Y on X*. If the roles of *X* and *Y* were reversed, and the values of *X* predicted from the variable labeled *Y*, then we would be looking at the regression of *X* on *Y*. In this second case, the errors in prediction are measured with respect to the *X* axis. Unless all the observed points fall on a 45-degree line, the two slopes are not equal. Thus the regression model is asymmetric—since the describing variable and the response variable are treated differently

and different fitted lines result, depending upon which variable the researcher decides is the response variable and which is the describing variable.

Note that the question of a possible causal relationship is not decided by calling one variable the describing variable and the other the response variable. The question of causality is a separate and often difficult issue. By effectively summarizing the data, the regression

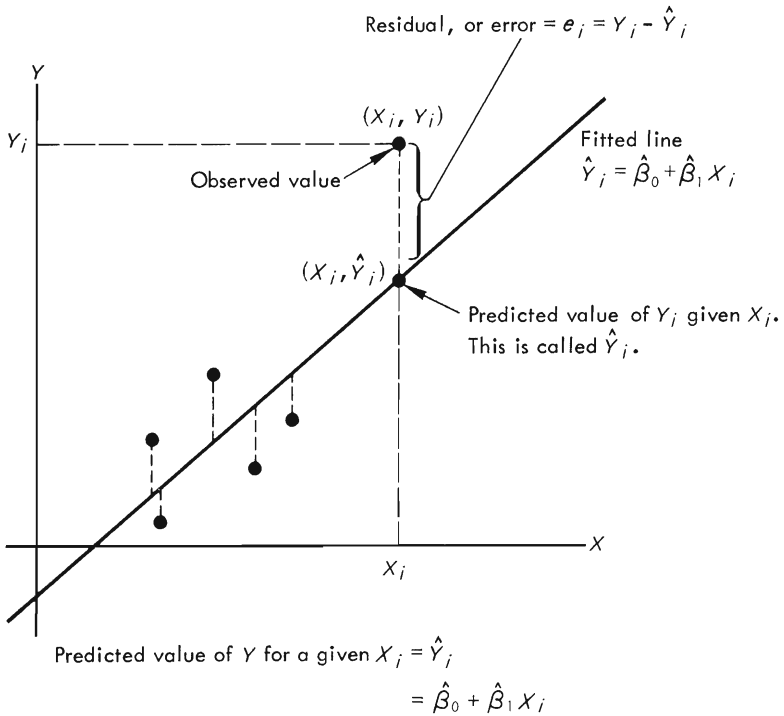


FIGURE 3-3 Notation for least-squares regression

analysis may sometimes provide some help in deciding if there is a causal relationship between the variables.

After fitting a line to a collection of data, the obvious question is: How well does the line fit? Here are four measures of the quality of fit:

1. the N residuals: $Y_i - \hat{Y}_i$,
2. the residual variation:

$$S_{Y/X}^2 = \frac{\sum (Y_i - \hat{Y}_i)^2}{N - 2},$$

3. the ratio of explained to total variation:

$$r^2 = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2},$$

4. the standard error of the estimate of the slope:

$$\frac{S_{Y|X}}{\sqrt{\sum (X_i - \bar{X})^2}}.$$

All these measures are functions of the residuals, $Y_i - \hat{Y}_i$. And all except the first are functions of the sum of squares of the residuals, $\sum (Y_i - \hat{Y}_i)^2$, which is the sum of squares minimized in estimating the parameters, β_0 and β_1 , of the fitted line. Such a functional dependence is not surprising, since reasonable measures of the quality of a line's fit to the data could hardly be anything except a function of the magnitude of the errors.

The residuals are particularly useful in assessing the fit of a line, since they are measured with respect to the Y axis—that is, they are measured in the same units as the response variable.

Instead of looking at the whole collection of N residuals—for there is a residual for each observation—we can summarize them by estimating the variability about the fitted line:

$$S_{Y|X}^2 = \frac{\sum (Y_i - \hat{Y}_i)^2}{N - 2}.$$

Sometimes the square root is taken, yielding the residual standard error for the fitted line.

Probably the most frequently used measure assessing the quality of fit of the line is r^2 , the proportion of the variance explained. Figure 3-4 shows the components of r^2 . For a given observation, $Y_i - \bar{Y}$ is the deviation of that observation from the mean, \bar{Y} . And $\sum (Y_i - \bar{Y})^2$ is the total variation in Y (that is, the sum of the squares of all the deviations from the mean). The describing variable seeks to predict or explain the individual deviations from the mean. The error in prediction for the i th observation is $Y_i - \hat{Y}_i$; and the error variation for all the observations is $\sum (Y_i - \hat{Y}_i)^2$. An intuitively sensible measure of the fit of the line is the ratio of this error or

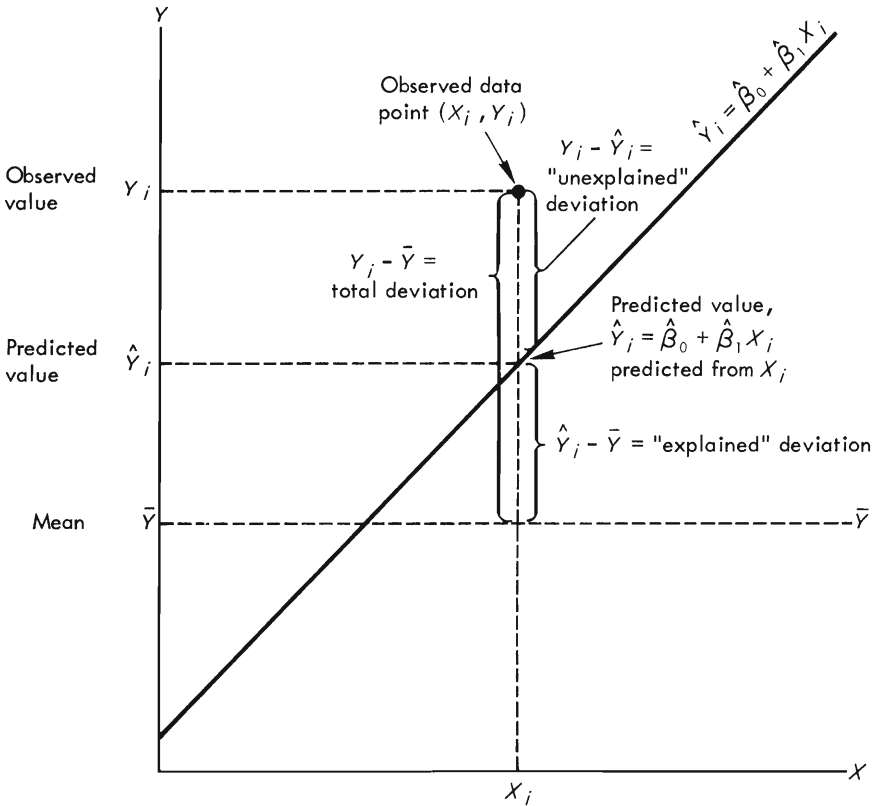


FIGURE 3-4 Components of r^2

unexplained variation to the total variation; the smaller this ratio, the better the fit:

one measure of fit

$$= \frac{\text{unexplained variation in } Y}{\text{total variation in } Y}$$

$$= \frac{\sum (Y_i - \hat{Y}_i)^2}{\sum (Y_i - \bar{Y})^2}$$

The commonly used measure, r^2 , is simply this ratio subtracted from one:

$$r^2 = 1 - \frac{\sum (Y_i - \hat{Y}_i)^2}{\sum (Y_i - \bar{Y})^2}$$

A little algebra proves that

$$\left(\begin{array}{c} \text{total} \\ \text{variation} \end{array} \right) = \left(\begin{array}{c} \text{explained} \\ \text{variation} \end{array} \right) + \left(\begin{array}{c} \text{unexplained} \\ \text{variation} \end{array} \right)$$

or

$$\Sigma (Y_i - \bar{Y})^2 = \Sigma (\hat{Y}_i - \bar{Y})^2 + \Sigma (Y_i - \hat{Y}_i)^2.$$

Therefore, since

$$r^2 = 1 - \frac{\text{unexplained variation}}{\text{total variation}}$$

we have

$$r^2 = \frac{\text{explained variation}}{\text{total variation}} = \frac{\Sigma (\hat{Y}_i - \bar{Y})^2}{\Sigma (Y_i - \bar{Y})^2}.$$

This interpretation of r^2 , as the ratio of explained to total variation, is very common. Often r^2 is expressed in percentage terms—for example, a value of r^2 of .51 will be described as “ X explained 51 percent of the variance in Y .” “Explained variance,” as used in the statistical jargon, refers only to the sum of squares, $\Sigma (\hat{Y}_i - \bar{Y})^2$. It may or may not refer to a good substantive explanation. A big r^2 means that X is relatively successful in predicting the value of Y —not necessarily that X causes Y or even that X is a meaningful explanation of Y . As you might imagine, some researchers, in presenting their results, tend to play on the ambiguity of the word “explain” in this context to avoid the risk of making an out-and-out assertion of causality while creating the appearance that something really was explained substantively as well as statistically.

If the fitted line has no errors of fit (that is, if the observed points all lie in a straight line), r^2 equals one, since there is no unexplained variation. At the other extreme, if the describing variable is no help at all in predicting the value of Y , r^2 will be near zero, since no variance is explained. In this unfortunate case, the regression line is simply $\hat{Y} = \bar{Y}$ (in other words, the predicted value of Y does not depend on the value of X).

In evaluating the fitted line, it is useful to know if the slope differs from zero. If the slope does not differ meaningfully from zero, then

X gives no help in explaining Y —the line is $\hat{Y} = \bar{Y}$. As explained in textbooks on statistics, a test of statistical significance and a confidence interval for the estimate of slope are constructed from the standard error of the estimate of the slope, which equals

$$S_{\hat{\beta}_1} = \frac{S_{Y|X}}{\sqrt{\sum (X_i - \bar{X})^2}}.$$

To conduct the test of statistical significance for $\hat{\beta}_1 \neq 0$, we consider the ratio of the estimated slope and its standard error:

$$\frac{\hat{\beta}_1 - 0}{S_{\hat{\beta}_1}}.$$

Under appropriate statistical assumptions, this has a t -distribution, with $N - 2$ degrees of freedom. For N greater than 30, the t -distribution closely matches the normal distribution. It is this match that gives rise to the rule of thumb that a regression coefficient should be roughly twice its standard error if it is to be statistically significant at the .05 level—since, for the normal, the two-tailed .05 limits are at ± 1.96 standard deviations.

Finally, note from the denominator of the formula for $S_{\hat{\beta}_1}$ that the error in the estimate of the slope grows smaller as the variability of X increases; that is, if the observations on the X variable are spread out instead of bunched together, the standard error of the estimate of the slope will be reduced. Consequently, if there is reason to believe that there is a linear relation between X and Y and if we can control the intervals at which X is measured, then it is better to choose values of X over a fairly wide range rather than bunched up together. For example, in a study of the effects of class size on teaching effectiveness, it would be better to construct classes of size 10, 15, and 20 students rather than 13, 15, and 17. By doing so, we might obtain a more secure estimate of the relationship between size and effectiveness.

This section has outlined the statistical mechanics of two-variable linear regression. We now apply the methods to a variety of data.

Example 1: Presidential Popularity and the Results of Congressional Elections

Let us, by way of review, apply all the different statistics estimated in the linear regression model to a single problem. Figure

3-5 shows the relationship between the President's approval rating (from the Gallup Poll) shortly before the midterm congressional election and the number of seats the President's political party loses in that congressional election, from 1946 to 1970. Table 3-1 shows

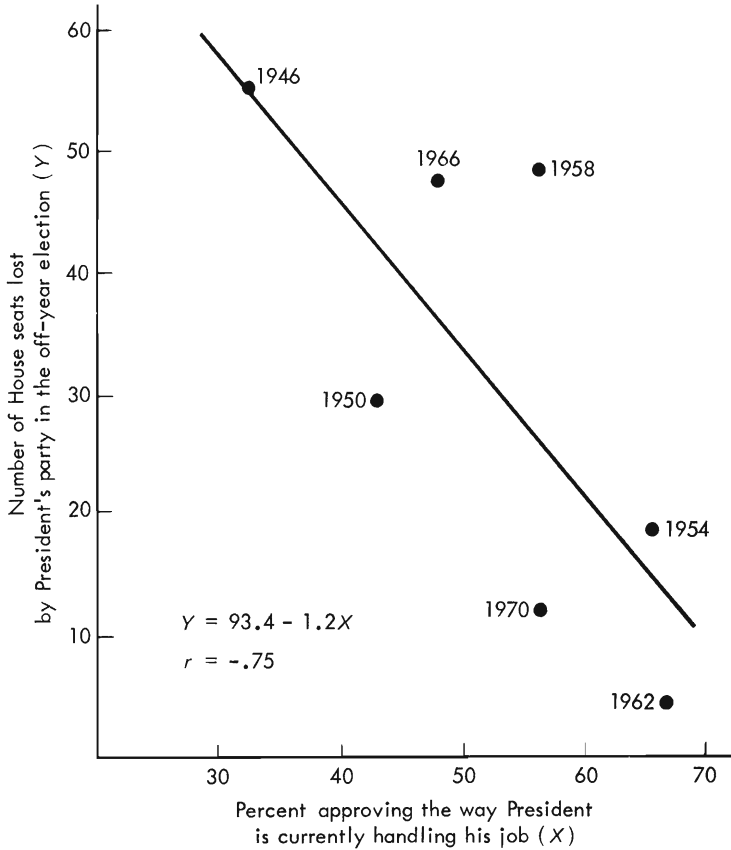


FIGURE 3-5 President's approval rating vs. his party's seat loss

the details of the data. Note that the political party of the President lost seats in each of the seven midterm elections from 1946 to 1970. Sometimes the loss was small—in 1962, for example, the Democrats lost only four seats in the House of Representatives compared to what they had in 1960. In other elections, many seats were lost: the Democrats suffered a decline of 55 Congressional seats in 1946. The Republicans, under President Eisenhower, had a bad year in the 1958 midterm elections, losing 48 seats.

Is, then, the extent of the loss of congressional seats by the President's

TABLE 3-1
Congressional Seats and Presidential Popularity

Year	Seats held in House of Representatives by		Seats lost in midterm election by President's party
	<i>Democrats</i>	<i>Republicans</i>	
1944	243	190	
1946	188	246	Democrats lost 55
1948	263	171	
1950	234	199	Democrats lost 29
1952	213	221	
1954	232	203	Republicans lost 18
1956	234	201	
1958	283	153	Republicans lost 48
1960	262	175	
1962	258	176	Democrats lost 4
1964	295	140	
1966	248	187	Democrats lost 47
1968	243	192	
1970	255	180	Republicans lost 12

Year	President's popularity rating early September in off-year elections (percent approve) ^a	
1946	Truman	32%
1950	Truman	43%
1954	Eisenhower	65%
1958	Eisenhower	56%
1962	Kennedy	67%
1966	Johnson	48%
1970	Nixon	56%

SOURCE: *Gallup Political Index*, October 1970, No. 64, page 16.

^aPercent approve + percent disapprove + percent no opinion = 100 percent. The question is worded as follows: "Do you approve or disapprove of the way Blank is handling his job as President?"

party related to the approval rating of the President?² The correlation between popularity and seat loss is, for the seven elections, $-.75$,

²Two papers dealing with the issues raised by these data are: Angus Campbell, "Voters and Elections: Past and Present," *Journal of Politics*, 26 (November 1964); 745-57, and John E. Mueller, "Presidential Popularity from Truman to Johnson," *American Political Science Review*, 64 (March 1970), 18-34. See also, for a more sophisticated discussion, Douglas A. Hibbs, Jr., "Problems of Statistical Estimation and Casual Inference in Dynamic, Time-Series Regression Models," in Herbert Costner, ed., *Sociological Methodology, 1973-1974* (San Francisco: Jossey-Bass, 1974), ch. 10.

indicating that the lower the President's popularity, the more seats his party loses in the off-year elections. This is, for most political research at least, a rather strong, impressive correlation—although note that the correlation coefficient doesn't tell us *how much* a decline in the approval rating is associated with a loss of *how many* seats. The regression coefficient does, however, provide some help with this. The equation of the least-squares line is

$$\text{seats lost} = 93.36 - 1.20 (\text{percent approving President})$$

Figure 3-5 shows this line. The slope is -1.20 , indicating that a one percent decline in the percent approving the current president is associated with a loss of about 1.2 seats in the upcoming off-year election. That regression coefficient is statistically significant:

$$t = \frac{\text{estimate of regression coefficient}}{\text{standard error}} = \frac{-1.20}{.48} = -2.50,$$

which, for five degrees of freedom, ($N - 2 = 7 - 2 = 5$) exceeds the one-tailed t -value at the .05 level (-2.02).

Furthermore, the President's approval rating explains a good deal of the statistical variation in the outcome of the election:

$$r = -.75, \quad r^2 = .56.$$

Thus the regression statistically explains 56 percent of the variation in the shifts in congressional seats.

All in all, this is a fairly impressive regression—a good correlation, a substantively meaningful regression coefficient that is statistically significant, and more than half the variance explained. Since it is so good, perhaps we can use the model for predictive purposes: taking the pre-election approval rating for the President and plugging into the regression equation to come up with an estimate of the loss of seats in the congressional election. This is all very nice, except that the prediction will not be a very secure one. Let us evaluate the quality of predictions based on the fitted line.

One way to get an idea of the predictive properties of the model is to look at the estimate of the variability about the line, the residual variance:

$$S^2_{Y|X} = \frac{\Sigma (Y_i - \hat{Y}_i)^2}{N - 2}.$$

The numerator is simply the unexplained variation. Taking the square root puts this statistic into the units in which the response variable, Y , is measured:

$$S_{Y|X} = 13.3 \text{ seats,}$$

which is a rather large standard error in terms of predicting seats—especially when we start to consider confidence intervals of \pm two standard errors.

Or, to evaluate the predictive quality of the model, we might look directly at the residuals for each year of the observed data. Table 3-2 shows the computations. Once again, we see pretty substantial errors in prediction from the observed data—and, of course, the model itself is estimated so as to minimize the sum of squares of these residuals.

In short, then, we have here the beginnings of a good explanatory model, but it still needs improvement if it is to be useful for predictive purposes. How might we build a better, more complete model? Consider a model that also takes into account the economic conditions—for which some voters might hold the President and his party responsible—prevailing at the time of the election:

$$\text{seats lost} = \beta_0 + \beta_1 (\text{presidential popularity}) + \beta_2 (\text{economic conditions}).$$

Just as in the two-variable case, this three-variable model is estimated by least squares. Such a multiple regression, as it is called, will be examined in Chapter 4.

TABLE 3-2
Residual Analysis

Year	$Y_i =$ observed seat loss by President's party	$X_i =$ Presidential approval rating	$\hat{Y}_i =$ predicted seat loss for a given X_i , $\hat{Y}_i =$ $93.4 - 1.20X_i$	Residual ^a = observed - predicted = $Y_i - \hat{Y}_i$
1946	55 seats	32%	$93.4 - 1.2(32) = 55$	$55 - 55 = 0$ seats
1950	29 seats	43%	$93.4 - 1.2(43) = 42$	$29 - 42 = -13$ seats
1954	18 seats	65%	$93.4 - 1.2(65) = 15$	$18 - 15 = 3$ seats
1958	48 seats	56%	$93.4 - 1.2(56) = 26$	$48 - 26 = 22$ seats
1962	4 seats	67%	$93.4 - 1.2(67) = 13$	$4 - 13 = -9$ seats
1966	47 seats	48%	$93.4 - 1.2(48) = 36$	$47 - 36 = 11$ seats
1970	12 seats	56%	$93.4 - 1.2(56) = 26$	$12 - 26 = -14$ seats

^aNote that if residual > 0 , the President's party lost more seats than predicted; if residual < 0 , the President's party lost less seats than predicted.

Example 2: Lung Cancer and Smoking

THE FITTED LINE

Figure 3-6 shows the relationship between the death rate from lung cancer in 1950 and the cigarette consumption in eleven countries in 1930. Cigarette consumption is lagged twenty years behind the death rate on the assumption that the carcinogenic consequences of smoking require a considerable length of time to show up. The fitted regression line is

$$\left[\begin{array}{l} \text{lung cancer deaths} \\ \text{per million people} \\ \text{in 1950 (Y)} \end{array} \right] = .23 \left[\begin{array}{l} \text{cigarettes consumed} \\ \text{in 1930 (X)} \end{array} \right] + 66,$$

$$\text{standard error of slope} = .07 \quad r^2 = .54$$

The regression indicates that when cigarette consumption in 1930 from one country to another is greater by, say, 500 cigarettes per year per person, the lung cancer rate apparently increased by about 115 deaths per million in 1950.

SCALING OF VARIABLES AND INTERPRETATION OF REGRESSION COEFFICIENTS

Note that in order to make an accurate interpretation of the regression coefficients, we must keep track of the units of measurement of each variable. For example, if the lung cancer rate were expressed as deaths per 100,000 people (instead of per 1,000,000), then the regression coefficient would be reduced by a corresponding factor of ten down to .023. This coefficient, although it is numerically smaller, reflects only the change in the scaling of the death rate—and the coefficient has exactly the same substantive meaning and importance as the original coefficient of .23. This obvious point is worth keeping in mind because some research reports are not particularly clear in reporting the units of measurement associated with each regression coefficient—and the reader must dig out the units of measurement and the scaling of the variables from the footnotes.

ANOTHER FITTED LINE: A REGRESSION WITHOUT THE UNITED STATES

A further look at the scatterplot shows the rather strong effect of one extreme point in shifting the fitted line. The line is pulled down

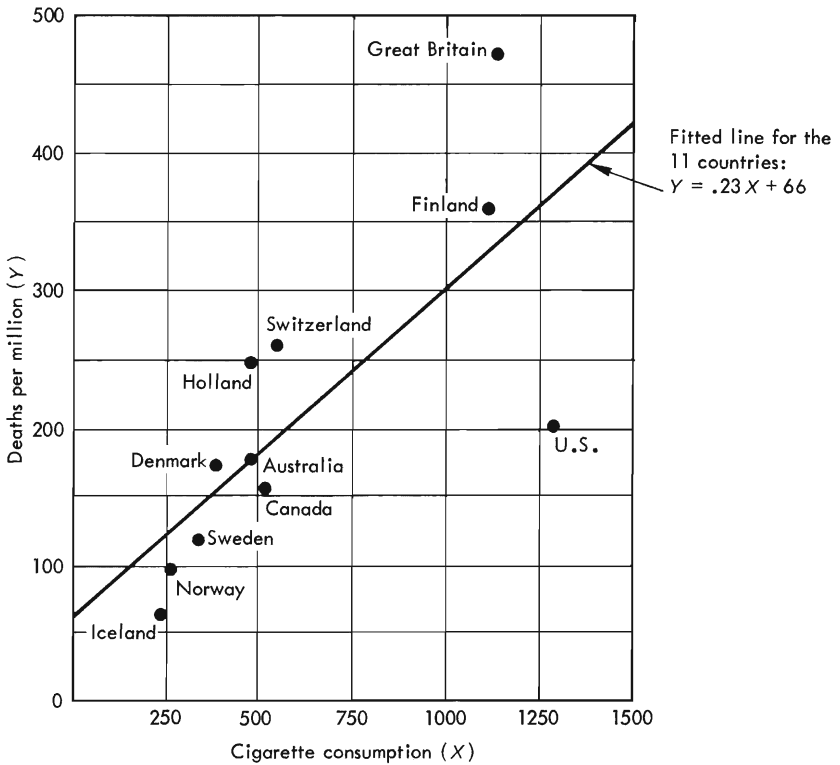


FIGURE 3-6 Crude male death rate for lung cancer in 1950 and per capita consumption of cigarettes in 1930 in various countries

SOURCE: R. Doll, "Etiology of Lung Cancer," *Advances in Cancer Research*, 3 (1955), reprinted in *Smoking and Health*, Report of the Advisory Committee to the Surgeon General (Washington: USGPO, 1964), p. 176.

by the low death rate for the United States. Removing that country from the data and computing a new regression line based on the remaining ten countries yields quite a different fitted line:

$N = 10$ Countries (Without U.S.)	$N = 11$ Countries (With U.S.)
$Y = .36X + 14$	$Y = .23X + 66$
$r^2 = .89$	$r^2 = .54$
Standard error of slope = .05	Standard error of slope = .07
Dotted line in Figure 3-7	Solid line in Figure 3-7

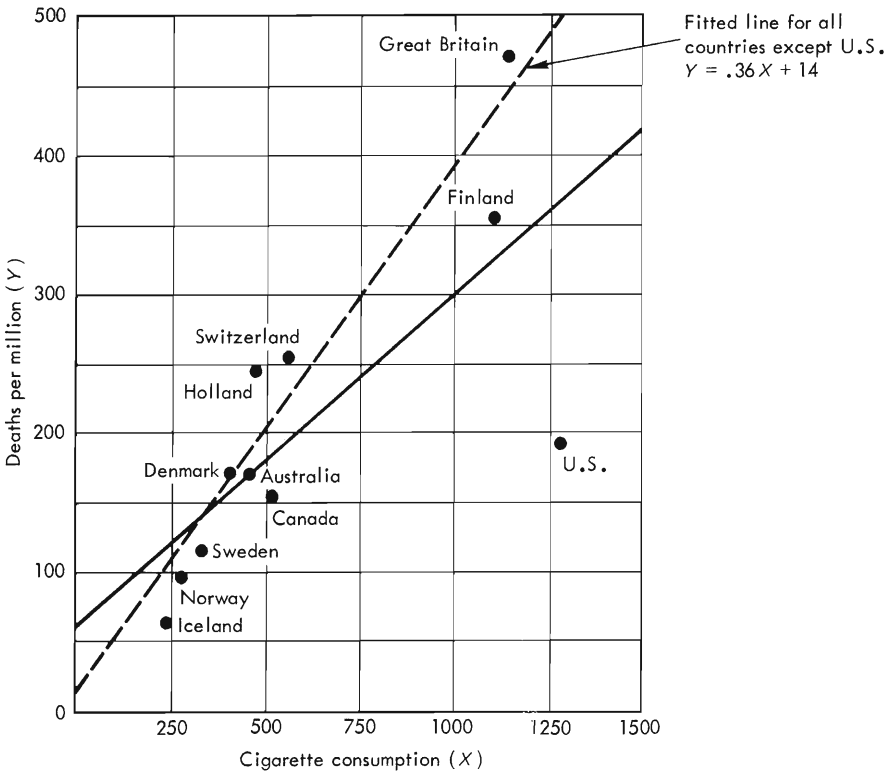


FIGURE 3-7 Lung cancer and cigarette consumption: fitted line for ten countries, omitting the United States

Note the great improvement in the explained variance in the regression based on the ten countries; a straight line really fits the ten quite well. Perhaps we should look more carefully into the conditions that make for a somewhat lower death rate than expected, given the amount of tobacco consumed, in the United States. That will be done below.

WHAT IF NOBODY SMOKED? INTERPRETING THE INTERCEPT

Let us return to consideration of the original regression for all eleven countries. Can we find out what the lung cancer rate might have been if there had been no smoking? Not very well with these particular data—for several reasons.

First, there is simply no experience at all with any countries consuming less tobacco per capita than Iceland, at 220 cigarettes per year per person in 1930. Obviously we want to be careful in

extending our results beyond the range of the data; some of the particular problems of extrapolation are discussed in Chapter 2.

Second, one naive way to answer the question meets some difficulties after a careful examination of the scatterplot. The naive approach is to set cigarette smoking at zero in the fitted regression equation and see what the lung cancer rate is. That rate is simply the intercept, 66 deaths per million per year. But note the pattern of countries down at the low end with respect to smoking: the three lowest countries have negative residuals, all lying below the fitted regression line. Thus, in the countries with a low consumption of cigarettes, there is some indication that a better-fitting curve would bend more sharply downward; thus the straight line imposed on the data is a bit misleading at the low end of the scale. This suggests that the rate would be considerably lower than 66 if nobody smoked. Perhaps a better estimate would be around 14 deaths per million—the intercept for the regression line that excluded the United States. The exclusion of that outlying value seems appropriate in estimating the intercept, since the outlier is far from the region of interest and since the residuals near the region of interest indicate that the extreme point has shifted the regression line based on all the countries.

Note finally that the line is literally imposed on the data—and just because we do the computations necessary to produce a slope and an r^2 , does not, of course, necessarily mean that the straight line is the best curve to fit to the data or that the two variables are, in fact, related in a linear fashion. In a later example, we will use “linear” regression to fit some other curves to data.

What kind of data *would* satisfactorily estimate the death rate from lung cancer if nobody smoked cigarettes? First, we need data based on individuals—smokers and nonsmokers—to make comparisons of lung cancer rates. Second, it is important to make sure that people susceptible—perhaps because of genetic or environmental factors—to lung cancer are not also people who are more likely to smoke. Thus we might compute the lung cancer rate for many different sorts of people who are smokers or nonsmokers. Such differential rates for different population groups could then be adjusted to the population as a whole to estimate the lung cancer rate if, contrary to fact, no one smoked.

ANALYZING THE RESIDUALS

Table 3-3 displays the original data, along with the predicted values for the lung cancer rate (predicted on the basis of cigarette consumption) and the errors made in the prediction for each country. Note

TABLE 3-3
Residual Analysis

Country	$Y_i =$ observed lung cancer deaths per million in 1950	$X_i =$ cigarettes consumed per capita in 1930	$\hat{Y}_i =$ predicted lung cancer death rate for a given X_i , $\hat{Y}_i = .23X_i + 66$	Residual = observed - predicted = $Y_i - \hat{Y}_i$
Iceland	58	220	.23(220) + 66 = 116	58 - 116 = -58
Norway	90	250	.23(250) + 66 = 123	90 - 123 = -33
Sweden	115	310	.23(310) + 66 = 137	115 - 137 = -22
Canada	150	510	.23(510) + 66 = 183	150 - 183 = -33
Denmark	165	380	.23(380) + 66 = 153	165 - 153 = 12
Australia	170	455	.23(455) + 66 = 170	170 - 170 = 0
United States	190	1280	.23(1280) + 66 = 359	190 - 359 = -169
Holland	245	460	.23(460) + 66 = 171	245 - 171 = 74
Switzerland	250	530	.23(530) + 66 = 187	250 - 187 = 63
Finland	350	1115	.23(1115) + 66 = 321	350 - 321 = 29
Great Britain	465	1145	.23(1145) + 66 = 328	465 - 328 = 137

the large residuals for Great Britain and the United States and the negative residuals for the smaller values of tobacco consumption. The residuals add up to zero; the sum of the squared residuals is the smallest it can be—no other line can improve over the least-squares line in minimizing the sum of the squares of the residuals. These two properties of the residuals—

- (1) $\Sigma (Y_i - \hat{Y}_i) = 0$, and
- (2) $\Sigma (Y_i - \hat{Y}_i)^2$ is minimized

—are properties of all least-squares lines.

A further analysis of the residuals can be made by plotting the residuals against the predicted values (\hat{Y}) as shown in Figure 3-8. Sometimes such a display yields up more information because the reference line is a horizontal line rather than the tilted line fitted to the original scatterplot. Contemplation of the residuals reveals large errors in the prediction of the death rate for Great Britain and the United States. Great Britain had a much higher death rate than the United States in 1950, although the per capita consumption of cigarettes in the two countries in 1930 was roughly equal. What differences between the two countries might account for the differences in lung cancer death rates even though the tobacco consumption was roughly the same? A few possibilities include:

1. Differences in air pollution between the two countries.

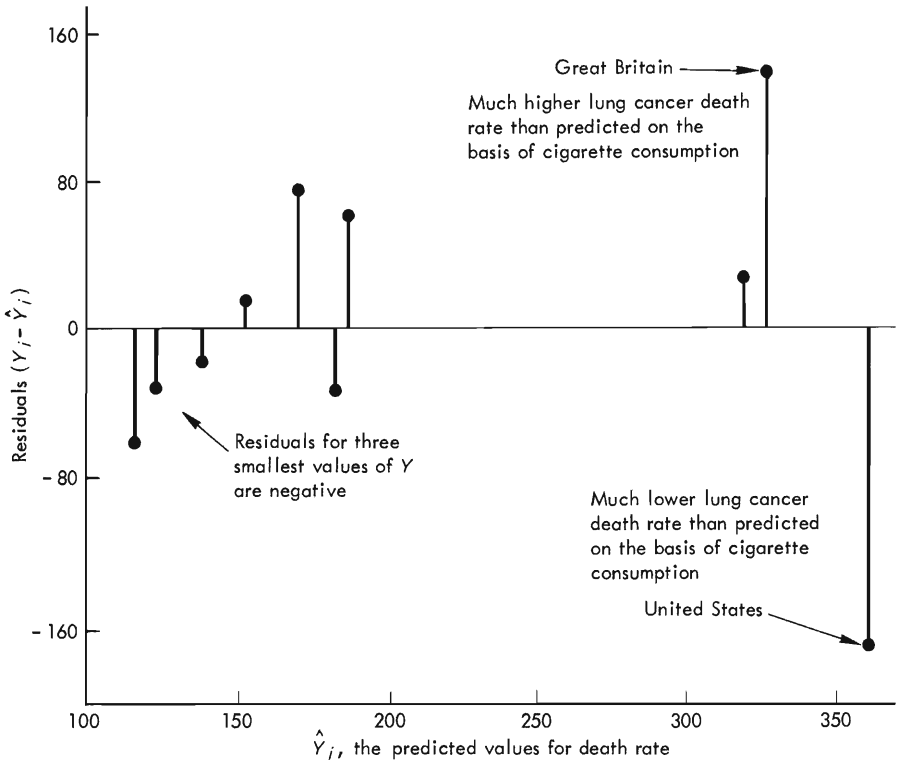


FIGURE 3-8 Residuals vs. predicted values, lung cancer and smoking

2. Differences in the age distribution of the populations of the two countries. Since lung cancer occurs more frequently among older smokers, the rate of cancer might well be higher in a country that had a larger share of older people.

3. Differences in smoking habits (such as smoking cigarettes right down to the end) that expose the lungs to different doses of smoke from each cigarette consumed. Observers have reported that the British often smoke their cigarettes right down to the very end (probably because cigarettes are heavily taxed and very expensive in England) and also that the British tend to be “drooper” smokers—they let the cigarette droop from the mouth rather than placing it in an ashtray or holding in the hand. Some researchers compared the lengths of discarded cigarette butts in the two countries and discovered rather large differences in length, the American discards being considerably

longer (30.9 mm) than the British (18.7 mm).³ Other studies found that “the mortality rate for lung cancer in England was especially high for the smokers who ‘drooped’ the cigarettes off the lip while they smoked, a habit which may result in the delivery of a greater dose of smoke from each cigarette.”⁴

4. Differences in the composition of the tobacco.

5. Differences in the factors which mute or accentuate the health consequences of smoking. For example, construction workers and others exposed to the insulating material asbestos who also smoke have a very high risk of lung ailments—a much higher risk than expected by merely adding up the excess risk from smoking plus the excess risk from working with asbestos. (This extra risk coming from the *combination* of the two factors is called, in the statistical jargon, an “interaction effect.”) Thus if more smokers in a country were exposed to asbestos, then that country would have a higher rate of lung cancer than expected on the basis of tobacco consumption alone.

6. Differences across countries in what medical symptoms doctors define or describe to be lung cancer.

VALUE OF THESE DATA AS EVIDENCE

These data have only a very modest value as evidence bearing on the relationship between smoking and lung cancer. Since the data are *aggregate, countrywide* figures, they provide very indirect evidence concerning the relationship between smoking and health among *individuals*. Furthermore, eleven data points aren’t much to work with—and the exclusion of a single observation shifted the variance explained from 54 percent to 89 percent, indicating the sensitivity of the analysis to outlying observations.

A big worry about the sort of data presented in Figures 3-6 and 3-7 is *selection*—how were the eleven countries included in the analysis chosen from all the countries of the world? Why these eleven? Would the results be the same if more countries were selected? Or eleven different countries? With so few data points, the analysis is very fragile; just a couple of fresh observations divergent from the fitted line would cause the whole relationship to fall apart. Careful, if manipulative, selection of data points can easily generate pseudo-rela-

³Report of the Advisory Committee to the Surgeon General of the Public Health Service, *Smoking and Health* (Washington, D.C.: U.S. Government Printing Office, 1959), p. 177.

⁴*The Health Consequences of Smoking, 1969 Supplement to the 1967 Public Health Service Review* (Washington, D.C.: U.S. Government Printing Office), p. 57.

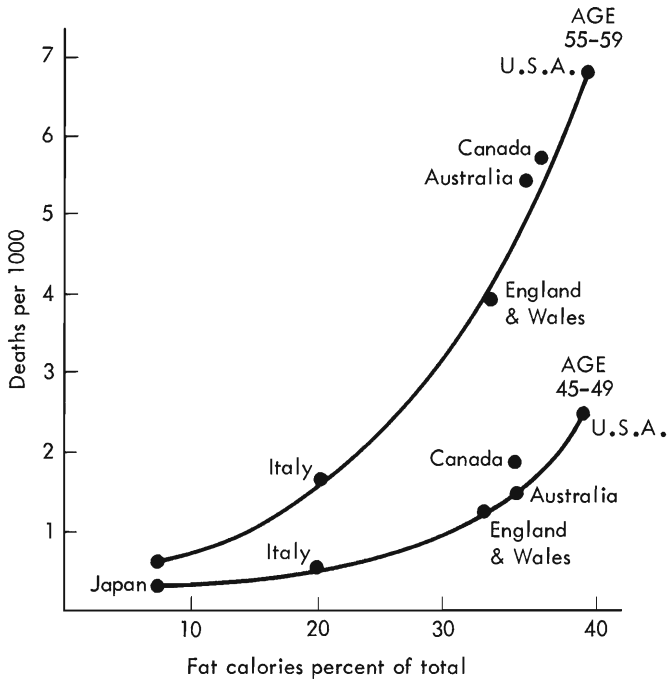


FIGURE 3-9 Mortality from degenerative heart disease (1948-1949, men) in relation to fat calories consumed
 SOURCES: Yerushalmy, *op. cit.* and Keys, *op. cit.* (see p. 87).

tionships. Yerushalmy points out such an example:

Another important error often encountered in the literature is the fallacy of utilizing evidence supporting a given hypothesis and neglecting evidence contradicting it. An illustration is shown in Figure [3-9]. In this case, the investigator selected six countries and correlated the percent of fat in the diet with the mortality of coronary heart disease in these six countries. . . . On the face of it, the correlation appears very striking, and indeed the author in reviewing the data in Figure [3-9] makes the following strong statement: "The analysis of international vital statistics shows a striking feature when the national food consumption statistics are studied in parallel. Then it appears that for men aged 40 to 60 or 70, that is, at the ages when the fatal results of atherosclerosis are most prominent, there is a remarkable relationship between the death rate from degenerative heart disease and the proportion of fat calories in the national diet. A regular progression exists from Japan through Italy, Sweden, England and Wales, Canada, and Australia to the United States. No other variable in the mode of life besides the fat calories in the diet is known which shows anything like such a consistent relationship to the mortality rate from coronary or degenerative heart disease."

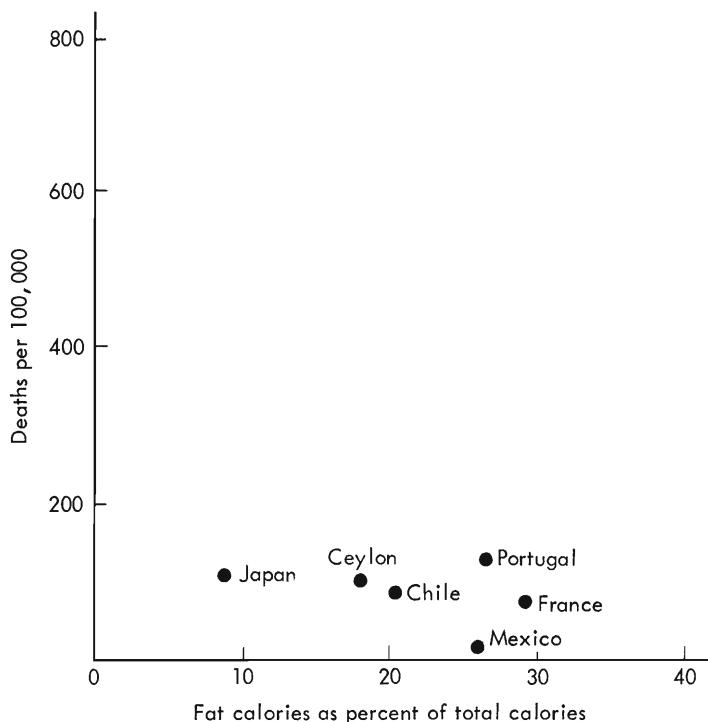


FIGURE 3-10 Six countries selected for equality in mortality from coronary heart disease, but differing greatly in consumption of fat calories in percent of total calories
 SOURCE: Yerushalmy, *op cit.* (see p. 87).

The question arises how were these six countries selected. Further investigation reveals that these six countries are not representative of all countries for which the data are available. For example, it is easy enough to select six other countries which differ greatly in their dietary fat consumptions, but have nearly equal death rates from coronary heart disease [Figure 3-10]. Similarly, six other countries were easily selected which consumed nearly equal proportions of dietary fat, but which differed widely in their death rates from coronary heart disease [Figure 3-11]. This tendency of selecting evidence biased for a favorable hypothesis is very common. For example, investigations among the Bantu in Africa are often mentioned in support of the dietary fat hypothesis of coronary heart disease, while observations on other African tribes, Eskimos, and other groups which do not support the hypothesis are generally ignored.

However, even when these errors are avoided and the studies are well conducted, the conclusions which may be derived from observational studies have great limitations stemming primarily from non-comparability of the self-formed groups. The phenomenon of self-selection is the root of many of the difficulties. Were all other

complications eliminated, the inequalities between groups which result from self-selection would still leave in doubt inferences on causality. For example, in the study of the relationship of cigarette smoking to health, if we assume well-conducted investigations in which (a) large random samples of the population have been selected and the individuals correctly identified as smokers, nonsmokers, or past smokers, (b) the problem of nonresponse did not exist, (c) the population had been followed long enough to identify all cases of the disease in question, (d) no problems of misdiagnosis and misclassification existed, (e) and no one in the population had been lost from observation, then even under these ideal conditions, the inferences that may be drawn from the study are limited because the individuals being observed, rather than the investigator, made for themselves the crucial choice: smoker, nonsmoker, or past smoker.⁵

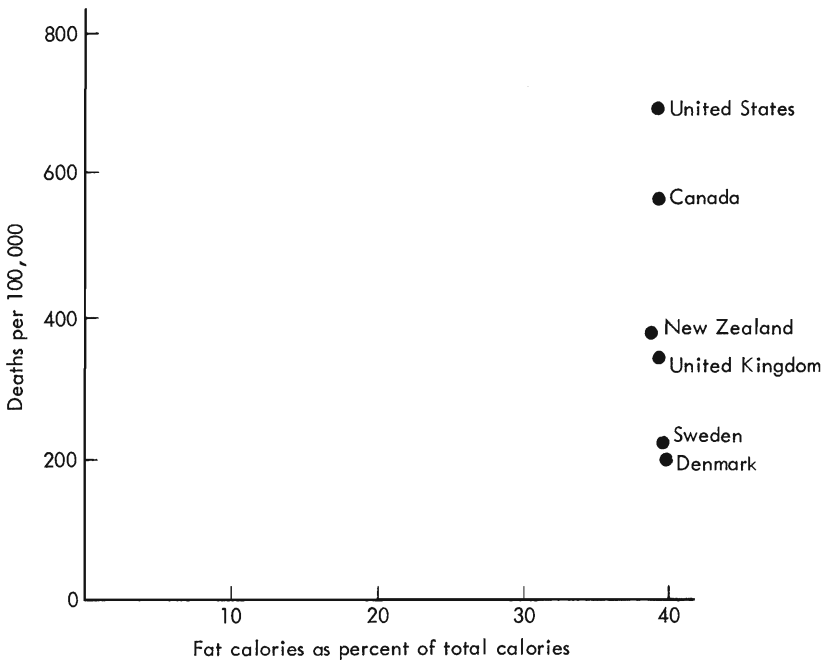


FIGURE 3-11 Six countries selected for equality in consumption of fat calories in percent of total calories, but differing greatly in mortality from coronary heart disease

SOURCE: Yerushalmy, *op. cit.*

⁵J. Yerushalmy, "Self-Selection—A Major Problem in Observational Studies," in Lucien M. Lecam, Jerzy Neyman, and Elizabeth L. Scott, eds., *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Biology and Health, Volume IV* (Berkeley and Los Angeles, California: University of California Press, 1972), pp. 332–33. The internal quotation is from A. Keys, "Atherosclerosis—A Problem in Newer Public Health," *Journal of Mt. Sinai Hospital*, 20 (1953), 134.

Still another reason for not taking our little analysis as serious evidence is that much better data are available to answer questions concerning the relationship between smoking and health. Smoking is probably the most carefully investigated public health problem there is; a vast amount of information has been gathered from health interviews with many people over many years, from autopsies, hospital records, animal studies, and so on. In other fields, where the amount and variety of evidence is less and the resources for collecting new data scarcer, the evidence of the sort examined here might represent the best available information and, furthermore, theories would have to stand or fall and decisions be made in the faint light of such analysis. Thus the overall importance of a particular piece of analysis varies in relation to what other evidence there is that bears on the question at hand.

Example 3: Increase in the Number of Radios and Increase in the Number of Mental Defectives, Great Britain, 1924–1937

The table shows a measure of the number of radios in the United Kingdom from 1924 to 1937 and the number of mental defectives per 10,000 people for the same years. These data form the basis for the discussion of “nonsense correlations” by the famous British statisticians, G. Udny Yule and M. G. Kendall.

The fit of the line is remarkably good, with a bit over 99% of the variation in number of mental defectives “explained” (in a statistical sense!) by the growth in the number of radios. Note the small, but systematic variation in the residuals, with the points weaving around the fitted line in clusters above and then below the fitted line. These “wrinkles” in the residuals might be worth pursuing if this were more than a nonsense correlation.

Why does this extremely strong, although nonsensical, relationship come about? This is a relationship formed by relating two increasing time series. In other words, the number of radios is increasing over time and also the number of mental defectives is increasing over time. Millions of other things are increasing over the time period from 1924 to 1937, including the population, the number of smokers, military expenditures in Europe, the number of patents issued, and the number of letters in the first name of the Presidents of the United States (Calvin, Herbert, and Franklin). For example, consider this regression:

Year	Number of radio receiver licenses issued (millions)	Number of notified mental defectives per 10,000 of estimated population
1924	1.350	8
1925	1.960	8
1926	2.270	9
1927	2.483	10
1928	2.730	11
1929	3.091	11
1930	3.647	12
1931	4.620	16
1932	5.497	18
1933	6.260	19
1934	7.012	20
1935	7.618	21
1936	8.131	22
1937	8.593	23

Figure 3-12 displays the regression line fitted to the above data:

$$\text{number of mental defectives per 10,000} = 2.20 \left[\begin{array}{l} \text{number of radios} \\ \text{(in millions)} \end{array} \right] + 4.58,$$

$$r^2 = .99, \quad \text{standard error of slope} = .08.$$

$$\text{number of mental defectives per 10,000 in the United Kingdom, 1924-1937} = 5.90 \left(\begin{array}{l} \text{number of letters} \\ \text{in the first name} \\ \text{of the President} \\ \text{of U.S., 1924-1937} \end{array} \right) - 26.44,$$

$$r^2 = .89, \quad \text{standard error of slope} = .66.$$

Yule and Kendall further observe:

. . . it might be argued that the period in question was one of great technical progress in many scientific fields; that one effect of this movement was the development of broadcasting and the general spread of the practice of listening evinced by the increased number of [radio] licenses taken out; that another effect was the greater interest in psychological ailments and increased facilities for treatment, resulting in either more discoveries of mental defect or greater readiness to submit cases to medical notice. Whether this is the right explanation is doubtful, but it is a possible rational explanation of what at first sight seems absurd.⁶

⁶G. Udny Yule and M. G. Kendall, *An Introduction to the Theory of Statistics* 14th ed., (London: Charles Griffin, 1950), p. 315-16.

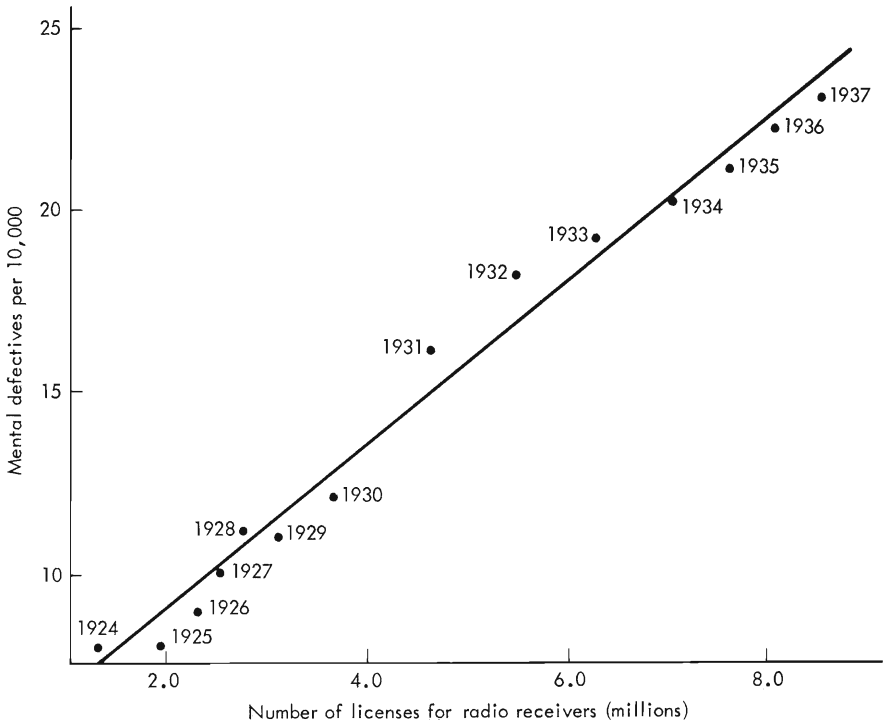


FIGURE 3-12 Radio receivers and mental defectives

Whether listening to the radio produced mental defectives (or, perhaps, whether the increase in number of mental defectives led to a greater demand for radios) is not answered by this regression of two increasing time series. And the relationship between the number of British mental defectives and the first names of American Presidents during 1924 to 1937 does not gain in credibility because the length of the name “explained” 87 percent of the variation in the number of mental defectives. What is clear, however, is that:

1. Even very high values of “explained” variance can occur without the slightest suspicion of a causal relationship between variables. There are times when a high value for r^2 might increase our degree of belief that there is a causal relationship, but this depends upon the substantive nature of the problem.
2. If nonsense goes into a statistical analysis, nonsense will come out. The nonsensical output will have all the statistical trappings, will look just as official, just as “scientific,” and just as “objective” as a substantively useful regression. It is, however, the substance and not the form that is the important thing. As Justice Holmes

once wrote: "The only use of forms is to present their contents, just as the only use of a pint pot is to present the beer . . . and infinite meditation upon the pot will never give you the beer."

We have now seen regression techniques applied to several problems—automobile safety inspections, smoking and lung cancer, and radios and mental problems. These examples all served to illustrate certain aspects of the logic and mechanics of fitting a line to the relationship between two variables. It is now time to examine a more extensive regression analysis in action, going into detail on a serious problem. Such is our next application.

Example 4: The Relationship between Seats and Votes in Two-Party Systems⁷

Arrangements for translating votes into legislative seats almost always work to benefit the party winning the largest share of the votes. That the politically rich get richer has infuriated the partisans of minority parties, encouraged those favoring majority parliamentary rule, and, finally, bemused a variety of statisticians and political scientists who have tried to develop parsimonious descriptions and explanations of the inflation of the legislative power of the victorious party. Here we will use a linear regression model to describe how the votes of citizens are aggregated into legislative seats and also to estimate the bias in an electoral system.

Figure 3-13 shows the data used in the analysis.⁸ These six scatter-plots indicate that the relationship between seats and votes in most two-party systems displays four obvious characteristics:

1. As a party's share of the vote increases, its share of the seats also increases in a fairly regular fashion.

⁷A more extended version of this material appeared in Edward R. Tufte, "The Relationship Between Seats and Votes in Two-Party Systems," *American Political Science Review*, 68 (June 1974), 540–54.

⁸The election tabulations were collected from state and national yearbooks. The U.S. congressional returns have been collected together in Donald Stokes and Gudmund Iversen, "National Totals of Votes Cast for Democratic and Republican Candidates for the U.S. House of Representatives, 1866–1960," July 1962, mimeo, Survey Research Center, University of Michigan. *Congressional Directories* (Washington, D.C.: U.S. Government Printing Office) were used to update the Stokes-Iversen compilation and also as the source for tabulations requiring election returns in individual congressional districts. All percentages of the vote were computed from the votes received by the two major parties only.

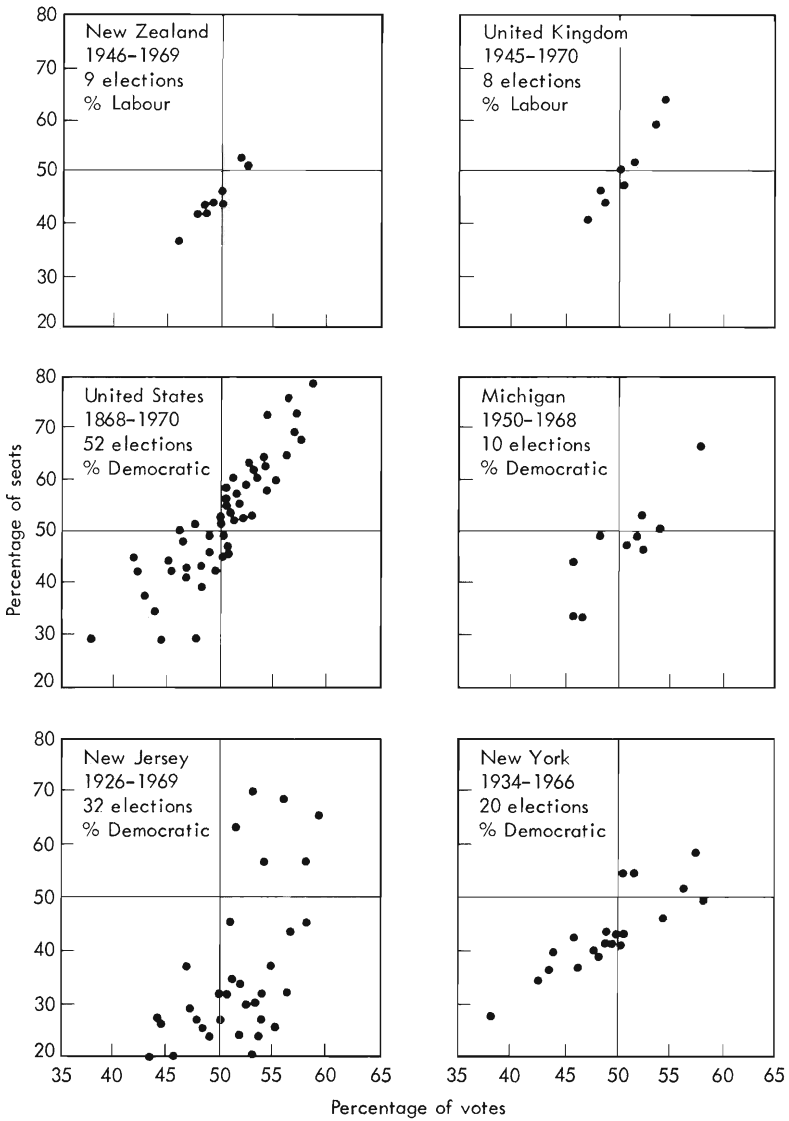


FIGURE 3-13 Seats and votes

2. The party that receives a majority of the votes usually receives a majority of parliamentary seats. Such was the case in 93 percent of the national elections and 53 percent of the state elections examined here. The points in the upper left and lower right quadrants represent those elections in which the party winning a majority of votes failed to take a majority of seats. New Jersey, like many other states prior to redistricting (and some after redistricting), shows many markedly biased outcomes, with the Democrats often winning fully three-fifths of the votes but less than one-third of the seats.
3. A party that wins a majority of votes generally wins an even larger majority of seats.
4. In most elections (100 percent in this series), the winning party receives less than 65 percent of the votes (although it may receive a much larger share of seats).

Even a casual inspection of the data displayed in Figure 3-13 indicates that almost any curve with a slope around two or three in the region from 35 to 65 percent of the vote for a party will fit the relationships rather well. Let us now examine the regression model.

The relationship between seats and votes is described most directly by a simple linear equation:

$$\left(\begin{array}{l} \text{percentage of seats for} \\ \text{a given political party} \end{array} \right) = \beta_1 \left(\begin{array}{l} \text{percentage of votes} \\ \text{for that party} \end{array} \right) + \beta_0.$$

The estimate of the slope, $\hat{\beta}_1$, measures the percentage change in seats corresponding to a change of one percent in the votes for a party. Thus $\hat{\beta}_1$ estimates the *swing ratio* or the *responsiveness* of the partisan composition of parliamentary bodies to changes in the partisan division of the vote in two-party systems. For example, the swing ratio during the last twelve U.S. congressional elections is 1.9, indicating that a net shift of 1.0 percent in the national vote for a party has typically been associated with a net shift of 1.9 percent in congressional seats for a party.

In addition, the fitted line provides an estimate of another important parameter of the electoral system: the bias for or against a particular party in the translation of votes into seats. Setting the percentage of seats at 50 percent and solving for the percentage of votes in the equation of the fitted line tells one the share of the vote that a party typically needs in order to win a majority of seats in the legislative body. The difference between this number and 50 percent is the *bias* or *party advantage*, as illustrated in Figure 3-14. For

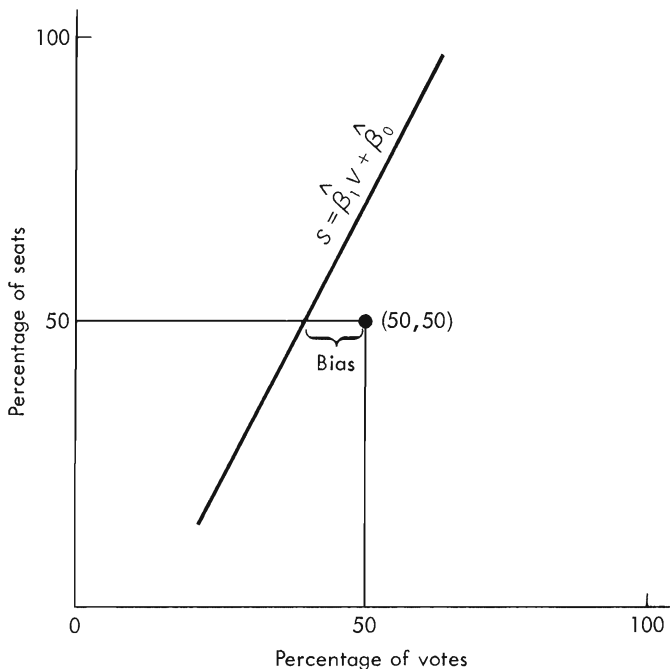


FIGURE 3-14 The fitted seats-votes line

example, in recent congressional elections, the Democrats have typically needed only about 48 percent of the national vote in order to win a majority of House seats; thus the bias or party advantage is about 2 percent. Later we will explain some of the variations in the swing ratio and bias for different electoral systems over the years.

Note that we are using the estimate of the slope in the linear model in order to estimate the swing ratio; the analogue of the intercept in the linear model is, in this case, the bias. Thus both the parameters estimated by the linear regression model are useful in this analysis.

One minor defect of the linear fit is that in general the fitted line will not pass through the end points (0 percent votes, 0 percent seats) and (100 percent votes, 100 percent seats), which are on the seats-votes curve by definition. Although slightly inelegant, this shortcoming is hardly troublesome—especially since parties in two-party systems almost never get less than 35 percent of the vote nor more than 65 percent of it.⁹ The clear advantage of the linear fit

⁹A “logit” model dealing with this problem is described in Example 6 of this chapter.

is that it yields two politically meaningful numbers, the swing ratio and the bias, that can be compared over time and electoral systems.

Table 3-4 records the fitted lines for a variety of elections. The swing ratios and the biases show considerable variation both between electoral systems and within some systems over time. Among the countries, Great Britain has the greatest swing ratio at 2.8. In the United States the swing ratio has been about two, although, as we shall see later, there is evidence that in the last few elections the swing ratio has decreased considerably. The U.K. electoral system shows little bias; in the United States a persistent bias has favored

TABLE 3-4
Linear Fit for the Relationship between Seats and Votes

	$\hat{\beta}_1$ Swing ratio and (standard error)	r^2	Percentage votes required to give the indicated party a majority of seats in the legislature	Advantaged party and amount of advantage
Great Britain, 1945-1970	2.83 (.29)	.94	50.2% Labour	Conservatives, .2%
New Zealand, 1946-1969	2.27 (.27)	.91	51.4% Labour	National, 1.4%
United States, 1868-1970	2.39 (.21)	.71	49.1% Democrats	Democrats, 0.9%
United States, 1900-1970	2.09 (.14)	.87	48.0% Democrats	Democrats, 2.0%
United States, 1948-1970	1.93 (.29)	.81	48.8% Democrats	Democrats, 1.2%
Michigan, 1950-1968	2.06 (.41)	.76	52.1% Democrats	Republicans, 2.1%
New Jersey, 1926-1947	2.10 (.44)	.53	61.3% Democrats	Republicans, 11.3%
New Jersey, 1947-1969	3.65 (.89)	.63	52.0% Democrats	Republicans, 2.0%
New York, 1934-1966	1.28 (.19)	.73	54.3% Democrats	Republicans, 4.3%

the Democratic party—partially the result of that party's victories in small congressional districts and in districts with low turnouts. In Michigan, New Jersey, and New York there have been large biases favoring the Republicans and a great deal of variation in swing ratios. The relationship between votes and seats is weaker for the three states than for the three countries; in fact, in the states during some time periods there was virtually no correlation between the share of seats that a party won in the legislature and the share of votes it had received at the polls! In more recent elections, however, there was a fairly strong relationship between seats and votes in all three states—probably the result of new rules and practices for districting.

THE SWING RATIO IN RECENT CONGRESSIONAL ELECTIONS

We now examine changes in the swing ratio in elections for the U.S. House of Representatives. Table 3-5 shows estimates of swing ratio and bias for congressional elections for the last hundred years. It appears that a shift—in fact, a rather striking shift—in the relationship between seats and votes has taken place in the last decade. The 1966–1970 triplet displays the second lowest swing ratio of the 17 election triplets since 1870. No doubt the recent elections provide a somewhat narrow range of electoral experience; the Democrats won with votes between 50.9 and 54.3 percent (a range in votes that is the fifth smallest of the 17 triplets). Until the Republicans control Congress or the Democrats win more decisively, the “new” swing ratio and bias will not be well estimated. The bias is a spectacular 7.9 percent, reflecting the two close votes that yielded the Democrats a substantial party majority in the House. The estimate of the bias for the 1966–1970 election triplet is, however, somewhat more insecure than for previous blocs of elections because the error of the estimated bias is proportional to the reciprocal of the swing ratio—and in this case the swing ratio is moderately small.

Compared with all the other performances of the electoral systems examined here, a system with a swing ratio of .7 and a bias of 7.9 percent describes a set of electoral arrangements that is both quite unresponsive to shifts in the preferences of voters (as expressed in their party votes for their representatives) and, at the same time, badly biased. How did the low value of the swing ratio for 1966–1970 come about? Certainly the Democratic party, after their substantial gain in votes (3.4 percent) and relatively tiny gain—given the “normal” swing ratio exceeding 2.0—in seats (3.2 percent) would like to know what happened in 1970. And for Republicans, 1966 and 1968 need

TABLE 3-5
Three Elections at a Time: Estimates of Swing Ratio and Bias

<i>Years of elections</i>	<i>Swing ratio</i>	<i>Percentage of votes to elect 50% seats for Democrats</i>	<i>Size of Democratic party advantage</i>
1870-74	6.01	51.4%	-1.4%
1876-80	1.48	50.0%	.0%
1882-86	3.30	50.8%	-.8%
1888-92	6.01	50.9%	-.9%
1894-98	2.82	51.7%	-1.7%
1900-04	2.23	50.1%	-.1%
1906-10	4.21	48.8%	1.2%
1912-16	2.39	48.8%	1.2%
1918-22	1.96	47.6%	2.4%
1924-28 ^a	-5.75 ^a	40.8% ^a	9.2% ^a
1930-34	2.28	45.9%	4.1%
1936-40	2.50	47.1%	2.9%
1942-46	1.90	48.1%	1.9%
1948-52	2.82	49.5%	.5%
1954-58	2.35	50.1%	-.1%
1960-64	1.65	47.4%	2.6%
1966-70	.71	42.1%	7.9%

^aThe figures estimated for the 1924-1928 election triplet are peculiar because of the extremely narrow range of variation in the share of the vote (42.1, 41.6, and 42.8 percent) during that period. The average range within an election triplet is about 6 percent.

explanation: after all, they managed to make the national division of the vote very close but in neither year were they able to win even 45 percent of the House seats.

The swing ratio indicates the potential for turnover in representation. The smaller the swing ratio, the less responsive the party distribution of seats is to shifts in the preferences of voters. The extreme case is a swing ratio near zero; such a flat seats-votes curve means that the distribution of seats does not change with the distribution of votes. Figure 3-15 shows the strong relationship between the swing ratio and the turnover in the House of Representatives for election triplets since 1870. Note the steady drift downward over the years in both the swing ratio and the turnover. Since 1948, the swing ratio has shifted from 2.8 to 2.4 to 1.7, and, most recently, to 0.7. Similarly the turnover in the House has declined, reflecting

the long-run decrease in the intensity of competition for congressional seats.¹⁰

One element in the job security of incumbents is their ability to exert significant control over the drawing of district boundaries; indeed, some recent redistricting laws have been described as the Incumbent Survival Acts of 1974. It is hardly surprising that legislators, like businessmen, collaborate with their nominal adversaries to eliminate

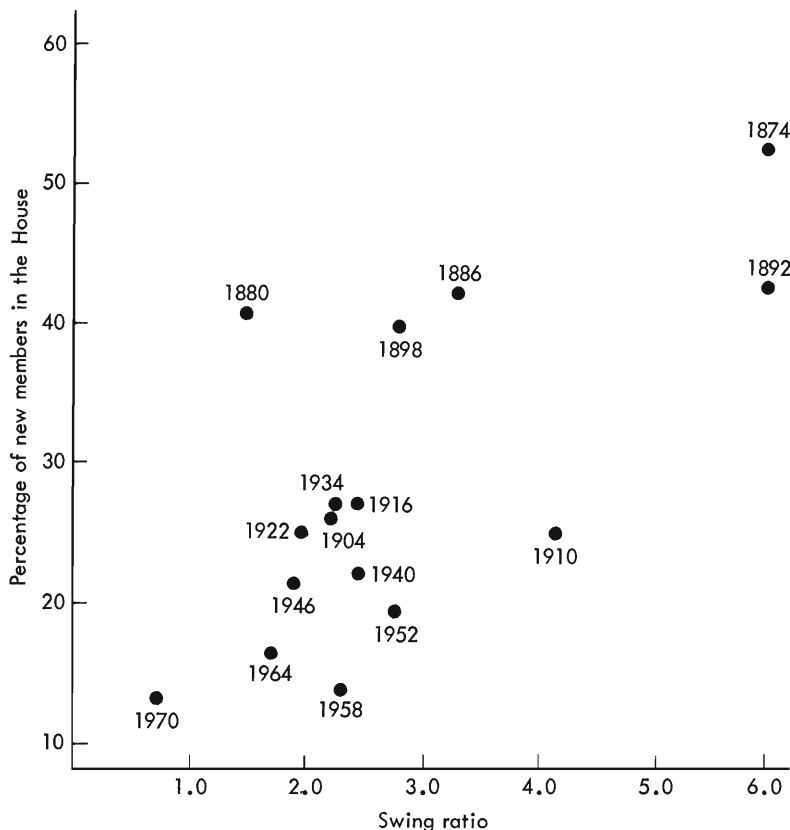
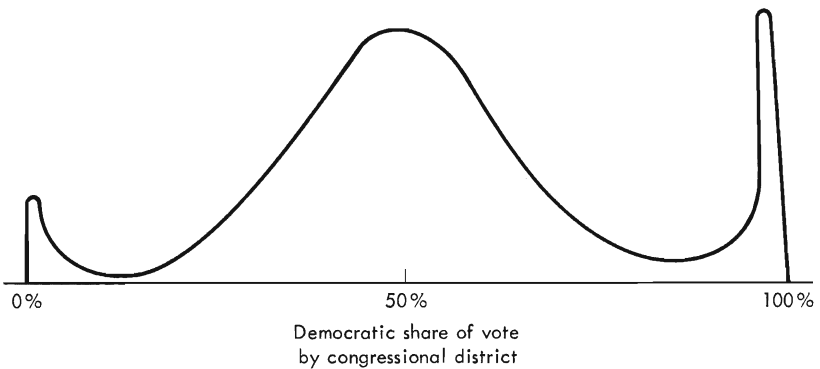


FIGURE 3-15 Turnover and swing ratio

dangerous competition. Ironically, reapportionment rulings have given incumbents new opportunities to construct secure districts for them-

¹⁰For example, Nelson W. Polsby, "The Institutionalization of the U.S. House of Representatives," *American Political Science Review*, 62 (March 1968), 144-68; and David R. Mayhew, "Congressional Representation: Theory and Practice in Drawing the Districts," in *Reapportionment in the 1970s*, ed. N. Polsby, pp. 249-90.

selves, leading to a reduction in turnover that is, in turn, reflected in the sharply reduced swing ratio of the last few elections. One apparent consequence is the remarkable change in the shape of the distribution of congressional votes in recent elections. Prior to 1964, the congressional vote by district was distributed the way everyone expects votes to be distributed: a big clump of relatively competitive districts in the middle, tailing off away from 50 percent with some peaks at the ends of the distribution for districts without an opposition candidate:



In recent elections the shape of the distribution of the vote by district has changed; Figure 3-16 shows the movement of district outcomes away from the danger area of 50 percent in recent years—note the development of bimodality in the 1968 and 1970 district vote compared to previous years (the left peak contains the Republican safe seats; the right peak contains the Democratic safe seats). Perhaps the best way to see how this pattern developed over time is to array the vote distributions over the years and riffle through them—like an old-time peep show—and watch the middle of the distribution sag and the areas of incumbent safety bulge in the more recent elections.

Many states, in part through recent reapportionments, have practically eliminated political competition for congressional seats—even compared to the relatively small proportion of competitive seats in the past. In the 1970 elections in Michigan, for example, not one of the 19 districts was a close contest; the *most* marginal Republican victor won 56 percent of the vote and the *most* marginal Democrat won fully 70% of the vote in his district. In Illinois, the most closely

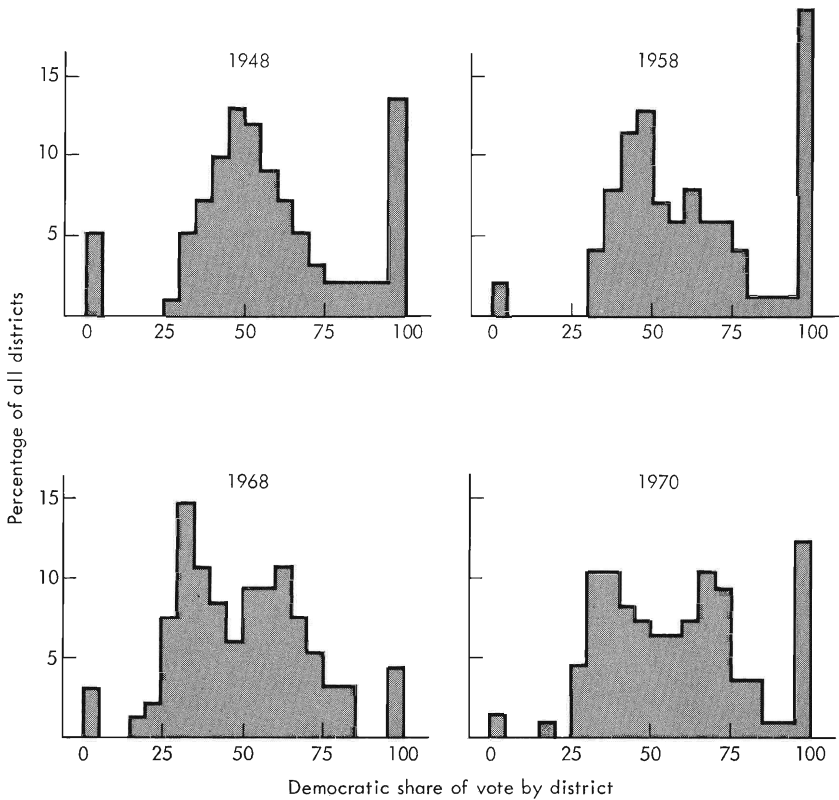


FIGURE 3-16 Distribution of congressional vote by district

contested race in all 24 congressional districts in 1970 was a 54–46 division of the vote; in contrast, in 1960, seven districts had closer races than that. The closest 1970 race in Pennsylvania was 55–45; in Ohio, 53–47.

In conclusion, then, we have seen here how the linear regression model can be used to measure two important qualities of an electoral system—the responsiveness and the partisan bias of the system. These two measurements might even be used by the courts to evaluate the fairness and the effectiveness of redistricting plans submitted to the courts.

This example has shown the economy of the regression model, in which the estimate of the slope takes us quickly to the central political issues in the data. There was little to learn from a correlation coefficient

in this case (and in many others), for we already knew that there was a strong relationship between how many votes and how many seats a party received. In contrast to the correlation coefficient, the regression model gave us a measure permitting politically meaningful comparisons across different political systems. Note also that a correlational analysis misses the method of assessing the partisan bias—an estimate which flows naturally from the regression model. Finally, look back at those four histograms in Figure 3-16. Note how informative they are with respect to the performance of the electoral system and how directly they make the point. Such is generally the case. Pictures of the data—charts, scatterplots, histograms, or just the values of a variable marked out on a line—are powerful aids to analysis. They also are easy to produce, either by hand or by computer.

Example 5: Comparing the Slope and the Correlation Coefficient

Both the correlation coefficient, r , and the slope of the fitted line, $\hat{\beta}_1$, are numerical summaries of the relationship between two variables. The slope, since it expresses the relationship in terms of the units in which X and Y are measured, is often a more useful summary measure than the correlation. This was true in the examples dealing with midterm congressional elections and the translation of votes into seats. In those examples the slope carried the important message in the data. Such interpretations of the slope require, however, that the units of measurement of the X and Y variables make some sort of interpretative sense.

For example, in examining responses to an interview questionnaire—and correlating relationships over the different responses to questions—it is difficult to interpret a measure of the rate of change on the intensity of feeling on one question with respect to the intensity of feeling on another. In such a case, the correlation coefficient may be more appropriate.

John Tukey has expressed these views strongly:

. . . [M]ost correlation coefficients should never be calculated. . . .
 [C]orrelation coefficients are justified in two and only two circumstances, when they are regression coefficients, or when measurement of one or both variables on a determinate scale is hopeless. . . .
 The other area in which correlation coefficients are prominent

includes psychometrics and educational testing in general. This is surely a situation where determinate scales are hopeless.¹¹

The correlation coefficient, r , can be interpreted in a number of ways. Its square, r^2 , is the proportion of variance in the response variable explained by the describing variable. Or it can be viewed as the average covariation of standardized variables:

$$r = \frac{1}{N} \sum_{i=1}^N \left(\frac{X_i - \bar{X}}{S_X} \right) \left(\frac{Y_i - \bar{Y}}{S_Y} \right).$$

That is, each observation is rescaled and measured in terms of how many standard deviations it is from the mean—for a given observation (X_i, Y_i) :

$$\frac{X_i - \bar{X}}{S_X} \quad \text{and} \quad \frac{Y_i - \bar{Y}}{S_Y}.$$

The product of the rescaled variables is averaged over all observations to yield the correlation coefficient.

Both the correlation coefficient and the slope can be dominated by a few extreme values in the data. Since we are working with products of deviations from the mean, a data point far from the mean on both variables can virtually determine the value of r and β_1 . Thus sometimes r and β_1 do not provide very good summaries of the relationship between X and Y . They fail when the relationship is nonlinear and when the data contain extreme outlying values.¹² The problems are easily detected from a scatterplot of the data. Thus one practical moral is that every calculation of r and β_1 should also involve an inspection of the scatterplot.

Let us now look at a series of scatterplots. First are examples in which the data are well described by the linear model: the data are

¹¹J. W. Tukey, "Causation, Regression, and Path Analysis," in O. Kempthorne, et al., eds., *Statistics and Mathematics in Biology* (Ames, Iowa: Iowa State College Press, 1956), pp. 38–39.

¹²In the case of many nonlinear scatterplots, the data can be transformed and the linear model estimated. Outliers can be treated by transformations, by removing them from the analysis, or by "Winsorizing" them (setting the most extreme value on a variable to the next most extreme). See Joseph B. Kruskal, "Special Problems of Statistical Analysis: Transformations of Data," *International Encyclopedia of the Social Sciences* (New York: Macmillan, 1968), vol. 15, 182–93; and F. J. Anscombe, "Outliers," *ibid.*, 178–82.

roughly oriented around a straight line with no extreme outliers (Figure 3-17).

We finally turn to some data sets for which the correlation and the fitted line fail to summarize the data effectively. Figure 3-18 shows three scatterplots with widely divergent patterns of relationship between X and Y . The first plot shows no relationship, discounting the one extreme outlier on both measures. The second plot suggests a moderately strong linear relationship between X and Y . The third plot reveals a rather marked curvilinear relationship between X and Y , revealing that as X increases, Y gets bigger even faster. Despite the great variation in the visual message, *the correlation between X and Y* is the same in all three cases. Also, the slopes do not differ greatly in the three cases.

Often a set of data for which the linear model is not immediately applicable can be transformed so the linear model is valuable. Or, to put it the other way around: many models with nonlinearities in the variables can be estimated by so-called "linear" regression.

For example, suppose we work with the logarithm of the one of the variables and have the model

$$Y = \beta_0 + \beta_1 \log X.$$

This model is estimated by letting $X' = \log X$ and then performing the usual least-squares regression for the model

$$Y = \beta_0 + \beta_1 X'.$$

Thus the criticism sometimes made that linear regression "assumes linearity" is a bit misleading, since the assumption can, in fact, be checked—and, if false, the model then redesigned for purposes of estimation. In fact, a better name for what this chapter has been all about is "fitting curves to relationships between two variables."

In summary, then, fitting lines to relationships between variables is often a useful and powerful method of summarizing a set of data. Regression analysis fits naturally with the development of causal explanations, simply because the research worker must, at a minimum, know what he or she is seeking to explain. The regression model is surprisingly flexible; and we now illustrate methods that increase its range of application.

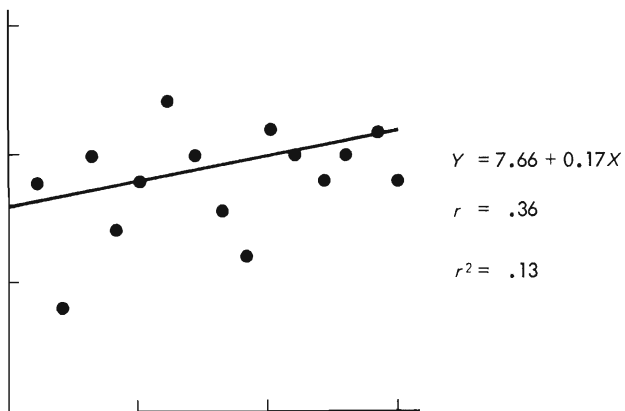
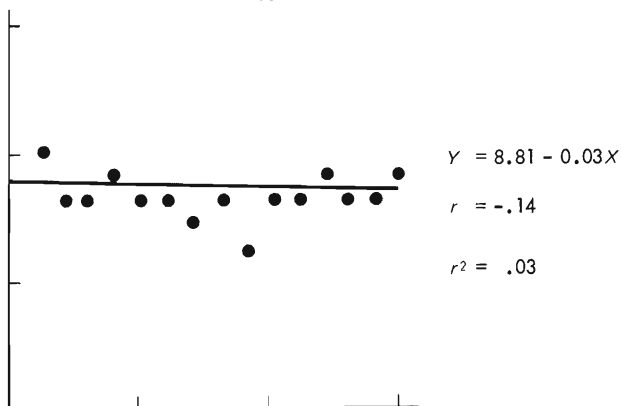
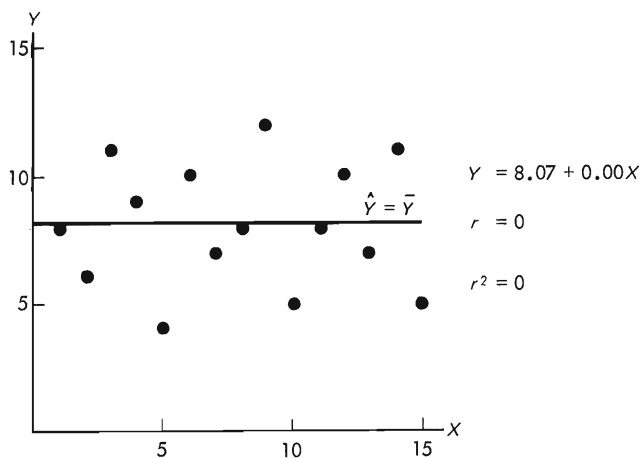


FIGURE 3-17 Data relatively well described by a fitted line

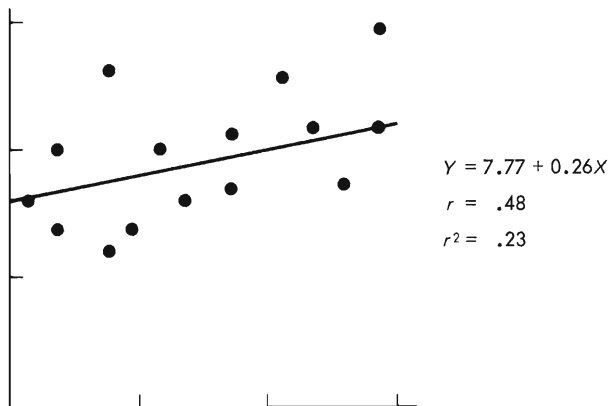
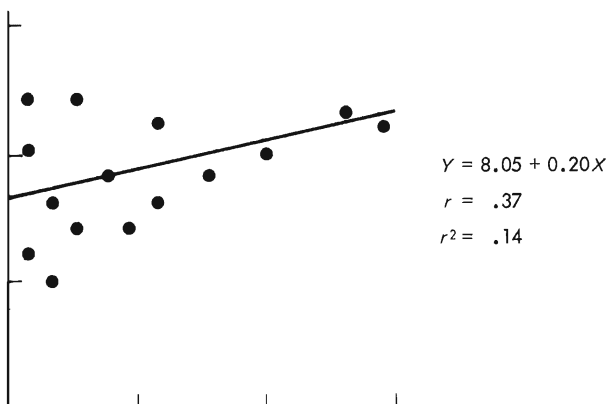
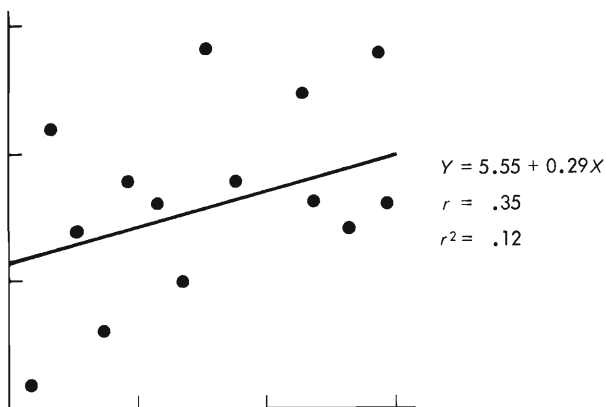


FIGURE 3-17 Data relatively well described by a fitted line

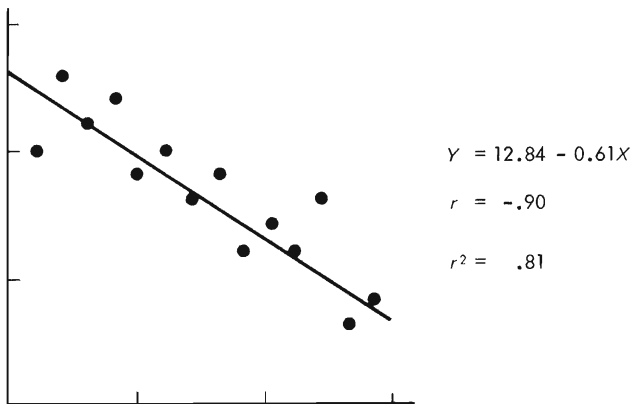
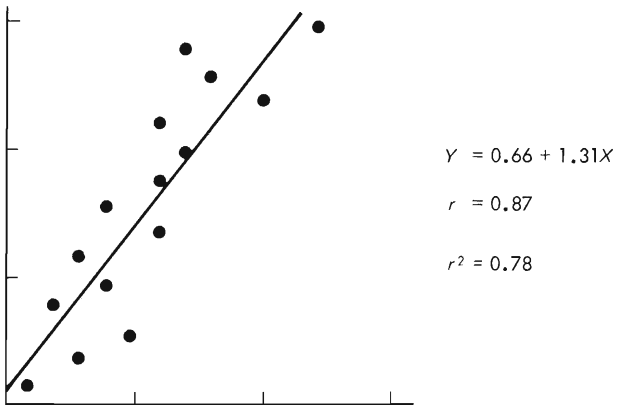
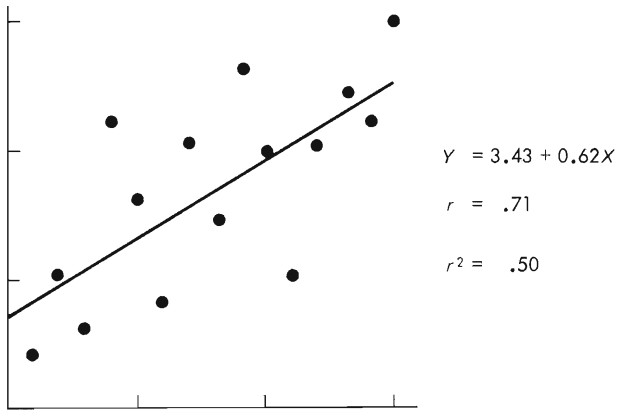


FIGURE 3-17 Data relatively well described by a fitted line

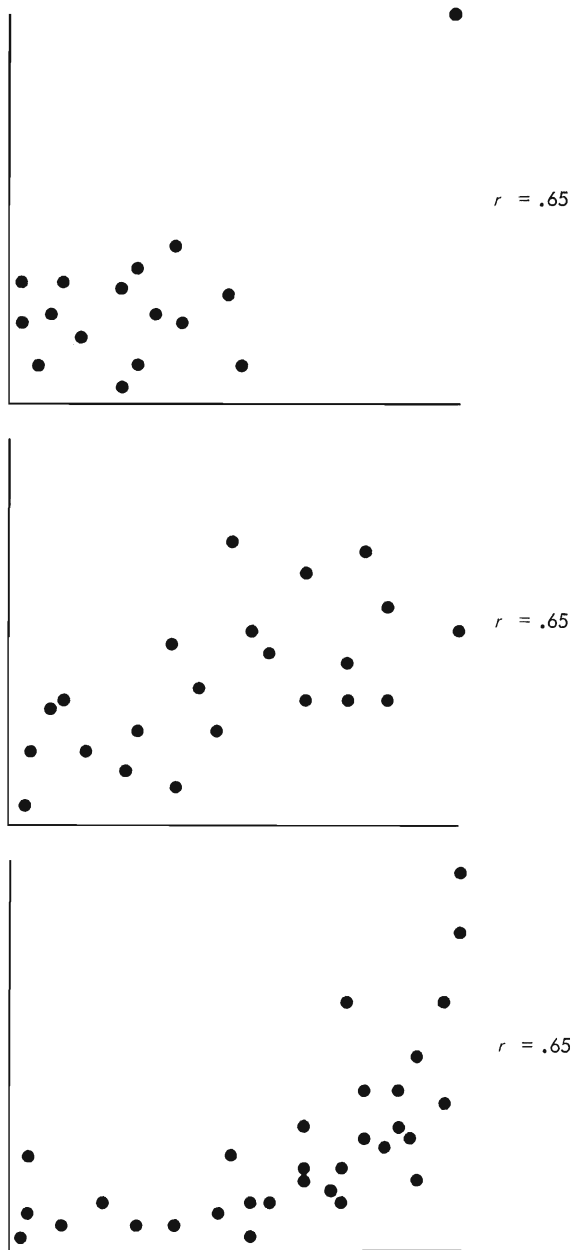


FIGURE 3-18 Three scatterplots with the same correlation

Example 6: Interpretation of Regression Coefficients when the Variables are Re-expressed as Logarithms (with Five Examples)

Data that are *counts* of populations, vital statistics, census data, and the like are almost always improved by taking logs. . . . Charles Winsor frequently prescribed the taking of logs of all naturally occurring counts (plus *one*, to handle that embarrassing quantity *zero*) before analyzing them—no matter what the sources [of the data].¹³

Often the logarithm of a variable is taken before entering that variable in a regression analysis. The logarithmic transformation serves several purposes:

1. The resulting regression coefficients sometimes have a more useful theoretical interpretation compared to a regression based on unlogged variables.
2. Badly skewed distributions—in which many of the observations are clustered together combined with a few outlying values on the scale of measurement—are transformed by taking the logarithm of the measurements so that the clustered values are spread out and the large values pulled in more toward the middle of the distribution.
3. Some of the assumptions underlying the regression model and the associated significance tests are better met when the logarithm of the measured variables is taken.

REMEMBERING LOGARITHMS

The logarithm to the base b of a number x , written as $\log_b x$, is the power to which the base must be raised to yield x . Thus

$$\log_{10} 1000 = 3, \text{ because } 10^3 = 1000.$$

Similarly:

$$\begin{array}{ll} \log_{10} 10,000 = 4, & \text{because } 10^4 = 10,000. \\ \log_{10} 1 = 0, & \text{because } 10^0 = 1. \\ \log_{10} 2 = .30103, & \text{because } 10^{.30103} = 2. \\ \log_{10} 2000 = 3.30103, & \text{because } 10^{3.30103} = 2000. \\ \log_{10} 20,000 = 4.30103, & \text{because } 10^{4.30103} = 20,000. \end{array}$$

In short, then, logarithms are powers of the base. The base 10, the base e (which forms what are called “natural” logarithms), and

¹³Forman S. Acton, *Analysis of Straight-Line Data* (New York: Wiley, 1959), p. 223.

the base 2 are the ones most commonly used. Logs to the base 2 take the following form:

$$\log_2 8 = 3, \text{ because } 2^3 = 8.$$

The logarithm of zero does not exist (regardless of the base) and therefore must be avoided. In logging variables with some zero values (especially those deriving from counts), the most common procedure is to add one to all the observations of the variable.

Finally, we should recall the following rules for manipulation of logarithms:

For $x > 0$ and $y > 0$:

$$\log xy = \log x + \log y.$$

For example,

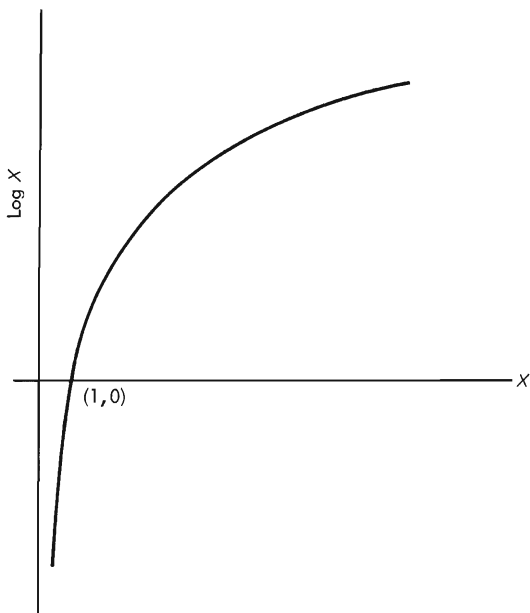
$$\begin{aligned} \log 20,000 &= \log (2)(10,000) \\ &= \log 2 + \log 10,000 \\ &= .30103 + 4 \\ &= 4.30103. \end{aligned}$$

$$\log \frac{x}{y} = \log x - \log y.$$

$$\log x^n = n \log x.$$

Let us first look at the effect of taking logarithms on the measurement scale of a single variable. Figure 3-19 shows the relationship between X and $\log X$; and Table 3-6 (page 111) tabulates the populations of some 29 countries of the world along with the logarithm of population. Note how the logarithmic transformation pulls the extremely large values in toward the middle of the scale and spreads the smaller values out in comparison to the original, unlogged values of the variable. Although the transformation preserves the rank ordering of the countries with respect to population, it still does produce quite a major change in the scaling of the variable here: the correlation between the population and the logarithm of population for the 29 countries is .68.

One reason for expressing population size here as a power of ten (that is, logging size to the base ten) is simply for convenience: if our scatterplots are going to include and differentiate between Iceland and Norway as well as the United States and India, then something must be done to compress the extreme end of the distribution. Logging

FIGURE 3-19 X vs. $\log X$

size transforms the original skewed distribution into a more symmetrical one by pulling in the long right tail of the distribution toward the mean. The short left tail is, in addition, stretched. The shift toward symmetrical distribution produced by the log transform is not, of course, merely for convenience. Symmetrical distributions, especially those that resemble the normal distribution, fulfill statistical assumptions that form the basis of statistical significance testing in the regression model. Figure 3-20 shows the contrast between the logged and unlogged frequency distributions of population.

Logging skewed variables also helps to reveal the patterns in the data. Figure 3-21 shows the relationship between the population size of a country and the size of its parliament—for the unlogged and the logged variables. Note how the rescaling of the variables by taking logarithms reduces the nonlinearity in the relationship and removes much of the clutter resulting from the skewed distributions on both variables; in short, the transformation helps clarify the relationship between the two variables. It also, as we will see now, leads to a theoretically meaningful regression coefficient.

Much of the value of the logarithmic transformation derives from its contribution to the testing of theoretical models by means of linear

TABLE 3-6
Population, 29 Countries, 1970

<i>Country</i>	<i>Population</i>	<i>Log (Population)</i>
Iceland	200,000	5.30
Luxembourg	400,000	5.60
Trinidad and Tobago	1,100,000	6.04
Costa Rica	1,800,000	6.25
Jamaica	2,000,000	6.30
New Zealand	2,800,000	6.45
Lebanon	2,800,000	6.45
Israel	2,900,000	6.46
Uruguay	2,900,000	6.46
Ireland	3,000,000	6.48
Norway	3,900,000	6.59
Finland	4,700,000	6.67
Denmark	4,900,000	6.69
Switzerland	6,300,000	6.80
Austria	7,400,000	6.87
Sweden	8,000,000	6.90
Belgium	9,700,000	6.99
Chile	9,800,000	6.99
Australia	12,500,000	7.10
Netherlands	13,000,000	7.11
Canada	21,400,000	7.33
Philippines	38,100,000	7.58
France	51,100,000	7.71
Italy	53,700,000	7.73
United Kingdom	56,000,000	7.75
West Germany	58,500,000	7.77
Japan	103,500,000	8.02
United States	204,600,000	8.31
India	554,600,000	8.74

regression.¹⁴ In interpreting regression coefficients of such models when the variables are logged, we have the following possibilities:

		Describing variable (<i>X</i>)	
		<i>Logged</i>	<i>Not logged</i>
Response variable (<i>Y</i>)	<i>Logged</i>	I	II
	<i>Not logged</i>	III	IV

¹⁴For further information see J. Johnston, *Econometric Methods*, 2d ed. (New York: McGraw-Hill, 1972), chap. 3; N. R. Draper and H. Smith, *Applied Regression*

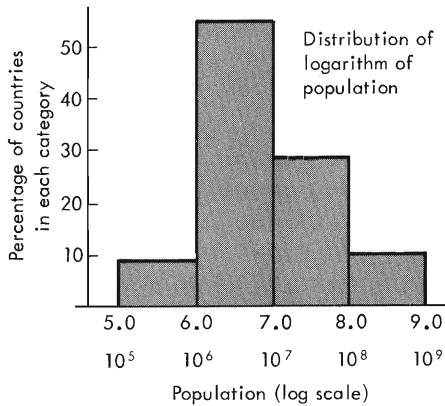
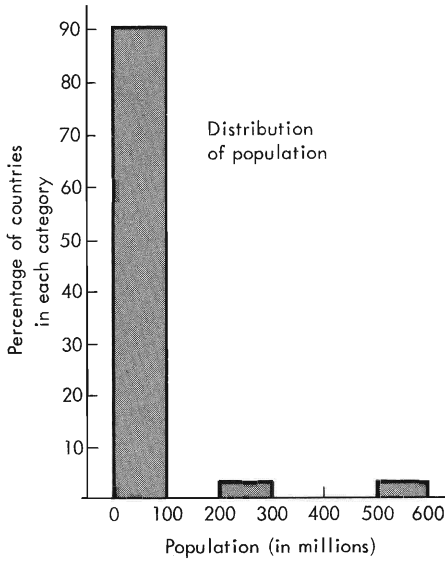


FIGURE 3-20 Logged vs. unlogged frequency distributions

Analysis (New York: Wiley, 1966); J. W. Richards, *Interpretation of Technical Data* (New York: Van Nostrand-Reinhold, 1967); and Joseph B. Kruskal, *op. cit.* For applications to political data see Hayward Alker and Bruce Russett, "Multifactor Explanations of Social Change," in Russett et al., *World Handbook of Political and Social Indicators* (New Haven, Conn.: Yale, 1964), 311-21.

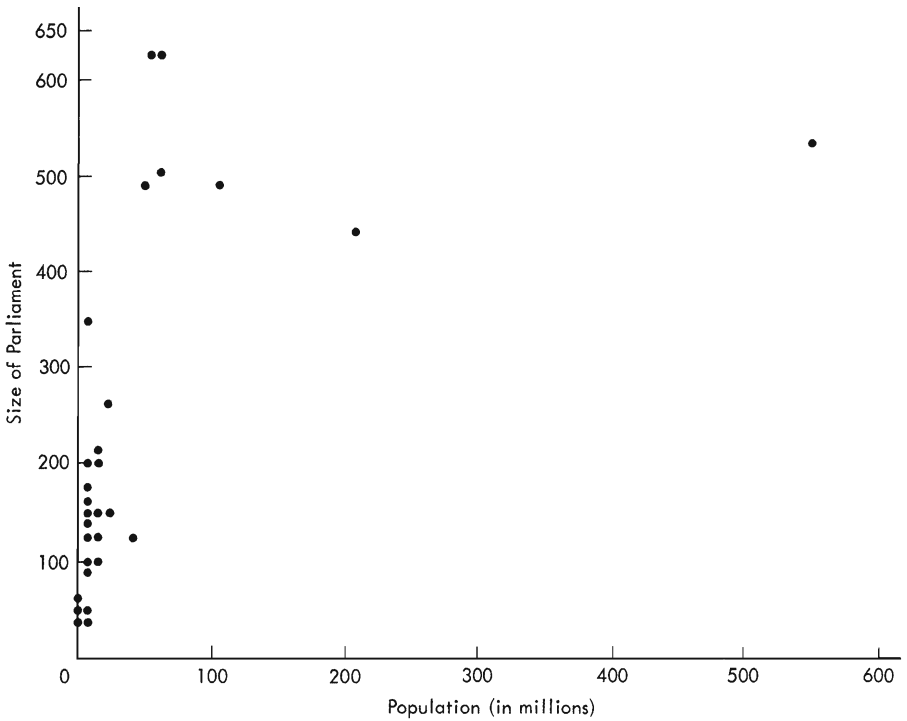


FIGURE 3-21a Relationship between parliamentary size and population, 29 democracies—neither variable logged

Case IV is simply the usual two-variable regression model with both variables unlogged. We now consider the three cases in which at least one of the variables in the analysis is logged.

CASE I—BOTH THE DESCRIBING AND THE RESPONSE VARIABLE LOGGED

In the model

$$\log Y = \beta_1 \log X + \beta_0,$$

we estimate β_1 and β_0 by ordinary least squares by letting $X' = \log X$ and $Y' = \log Y$, which yields the linear form

$$Y' = \beta_1 X' + \beta_0.$$

How is the regression coefficient in the double-log case interpreted? Beginning with the regression

$$\log_{10} Y = \beta_1 \log_{10} X + \beta_0$$

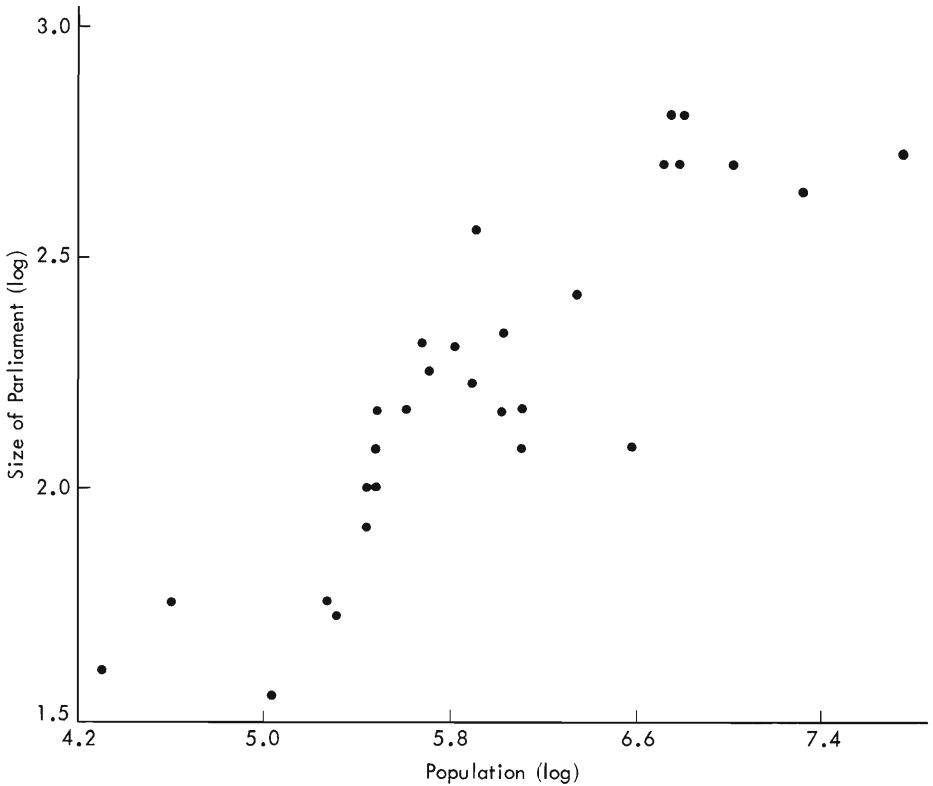


FIGURE 3-21b Relationship between parliamentary size (log) and population (log)—both variables logged.

and taking derivatives,

$$\frac{dY}{dX} \frac{1}{Y} \log_e 10 = \beta_1 (\log_e 10) \frac{1}{X} + 0,$$

yields
$$\frac{dY}{dX} \frac{X}{Y} = \beta_1$$

or
$$\beta_1 = \frac{dY/Y}{dX/X},$$
 which is the *elasticity* of Y with respect to X .

Thus β_1 measures the *percentage* change in Y with respect to a *percentage* change in X . The slope can be written approximately as

$$\beta_1 = \frac{\Delta Y/Y}{\Delta X/X}$$

and, when both the describing and the response variables are logged, the estimate of the slope assesses the proportionate change in Y resulting from a proportionate change in X . Note how this differs from the usual interpretation of the slope when both variables are unlogged (case IV):

$$\beta_1 = \frac{\Delta Y}{\Delta X}.$$

It is important to realize that fitting the model

$$\log Y = \beta_1 \log X + \beta_0,$$

does not *test* the assumption that there is, in fact, a proportionate relationship between X and Y . The logic is: *Assuming that there is a proportionate relationship between X and Y* , what is the best estimate of that proportionality or elasticity? Thus the regression answers the quantitative question by estimating a parameter in a model—on the assumption that the model is correct. We choose between competing models by comparing their goodness of fit, by thinking about their theoretical underpinnings, and by adding sufficient degrees of freedom in the model to allow the data to indicate the best fit. Our first example illustrates this point.

EXAMPLE 1 FOR THE LOG-LOG CASE: RELATIONSHIP BETWEEN PARLIAMENTARY SIZE AND POPULATION SIZE

Figure 3-22 shows the relationship, with both variables logged, between the population of a country and the size of its parliament for 135 countries of the world.¹⁵ This relationship appears nearly linear in logarithms, and the fitted line is

$$\log_{10} \text{ members} = .396 \log_{10} \text{ population} - .564,$$

which explains, statistically at least, some 70.7 percent of the variation

¹⁵A discussion of the substantive issues involved in this relationship is found in Robert A. Dahl and Edward R. Tufte, *Size and Democracy* (Stanford, Calif.: Stanford University Press, 1973), Ch. 7.

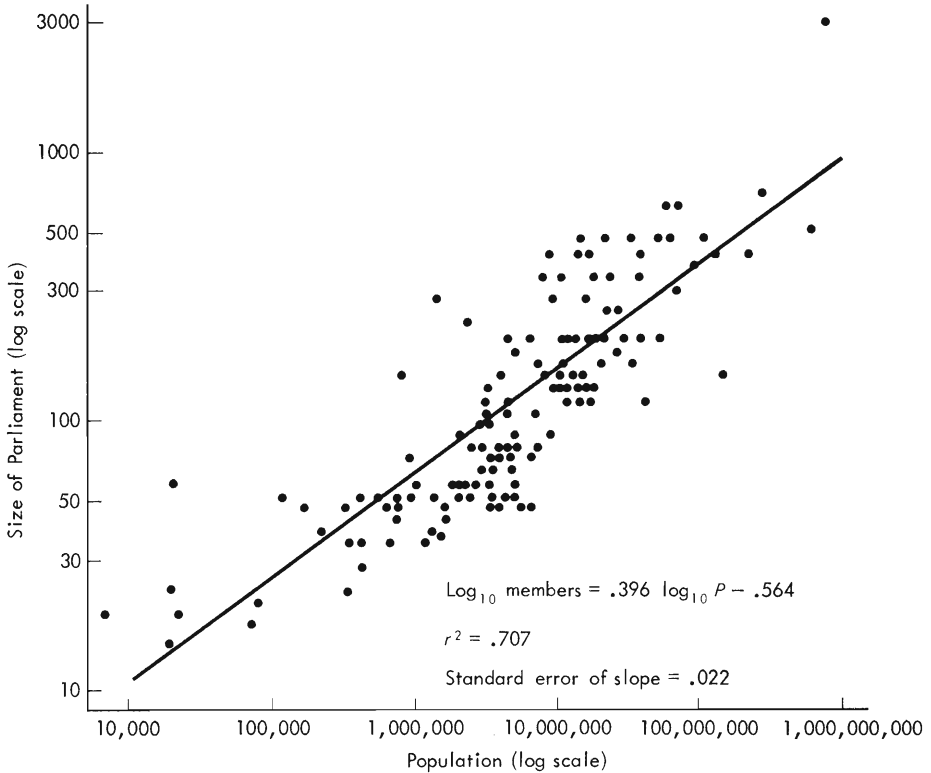


FIGURE 3-22 Population vs. parliament size—both variables logged

in parliamentary size. The estimated slope, .396, indicates that if a country was one percent above the average population of all countries, it was also typically about .4 percent above average with respect to size of parliament. A slightly more daring interpretation is to say that a change of one percent in population typically produces a change of .4 percent in parliamentary size.

Figure 3-22 and the residuals from the fitted line show a bend in the data—there is something of a threshold in the size of parliament for the smaller countries. For most of the countries with less than one million people, the observed points lie above the fitted line, indicating a tendency toward a minimum size of parliament around thirty members. We can improve upon the first fitted line for the 135 countries by examining some models that avoid the assumption of constant elasticity for all values of population (P) and take the bend in the data into account. One good approach, upon observing

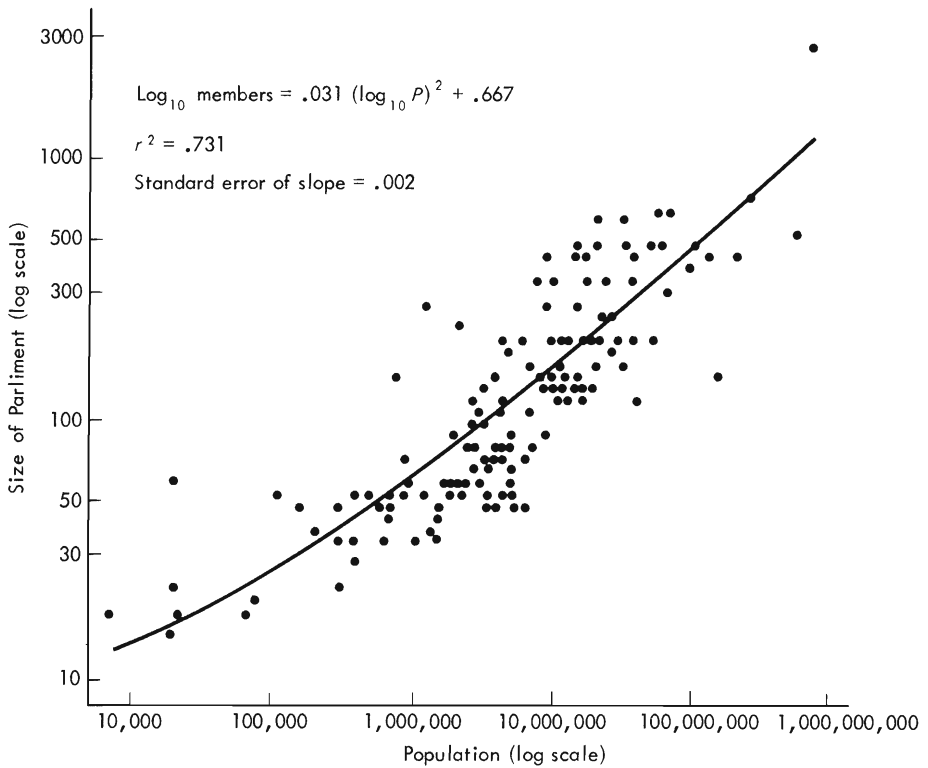


FIGURE 3-23 Fitted line with quadratic term

a curve in the data, is to introduce a quadratic term. The following fit, with its $(\log P)^2$ term, is our second model:

$$\log M = .031(\log P)^2 + .667.$$

Figure 3-23 shows the fit. This regression predicts 73.1 percent of the variation in the logarithm of parliamentary size—an improvement of 2.4 percentage points over the first model with no increase in the number of coefficients used in the model. What is the interpretation of this result? In particular, what does the regression coefficient mean? We get the answer by applying the same logic used in deriving the elasticity in the log-log case. The model is

$$\log_{10} M = \beta_0 + \beta_1 (\log_{10} P)^2.$$

Taking derivatives, as before,

$$\frac{dM}{dP} \frac{1}{M} \log_e 10 = 2\beta_1 (\log_e 10)(\log_{10} P) \frac{1}{P},$$

which yields

$$\begin{aligned} \frac{dM}{dP} \frac{P}{M} &= \text{elasticity of } M \text{ with respect to } P \\ &= 2\beta_1 \log_{10} P, \end{aligned}$$

or, in our particular case,

$$= .062 \log_{10} P.$$

Thus in this model the elasticity of M with respect to P is a slowly increasing function of $\log P$. For countries around 100,000, the elasticity of parliamentary size with respect to population is about .3; for countries of 100,000,000, it is nearly .5. Table 3-7 tabulates the relationship.

TABLE 3-7
Predictions of the Second Model

<i>Population</i>	<i>Log population</i>	<i>Elasticity of M with respect to P = .062 log₁₀ P</i>
10,000	4	.248
100,000	5	.310
1,000,000	6	.372
10,000,000	7	.434
100,000,000	8	.496
750,000,000	8.875	.550

The first model assumes that the elasticity is constant and provides an estimate under that untested assumption. The second model assumes that the elasticity varies as the population varies and provides an estimate under that untested assumption. The second is now favored because (1) visual inspection of the scatterplot and the residuals shows a bend in the data and (2) the second explains more variance than the first, even though both models estimate the same number of coefficients.

EXAMPLE 2 FOR THE LOG-LOG CASE: SIZE OF GOVERNMENTAL BUREAUCRACY AND POPULATION SIZE

For the fifty U.S. states, let B = the number of employees of the state government and let P = the number of people living in the state. Both P and B are highly skewed variables, and so we will work with $\log P$ and $\log B$. Figure 3-24 shows $\log B$ plotted against $\log P$.

Three sorts of general results could emerge from this analysis: (1) if a kind of Parkinson's Law held, then we would expect the bureaucracies of state governments to grow faster than the size of the state; (2) if there were, say, economies of scale, then we would expect bureaucracies to grow more slowly than the population of the state; and (3) the number of bureaucrats could grow in constant proportion to the size of the state. Obviously, other sorts of explanations can be used to explain the results of the analysis. The point here is that the number of employees of the state government can grow

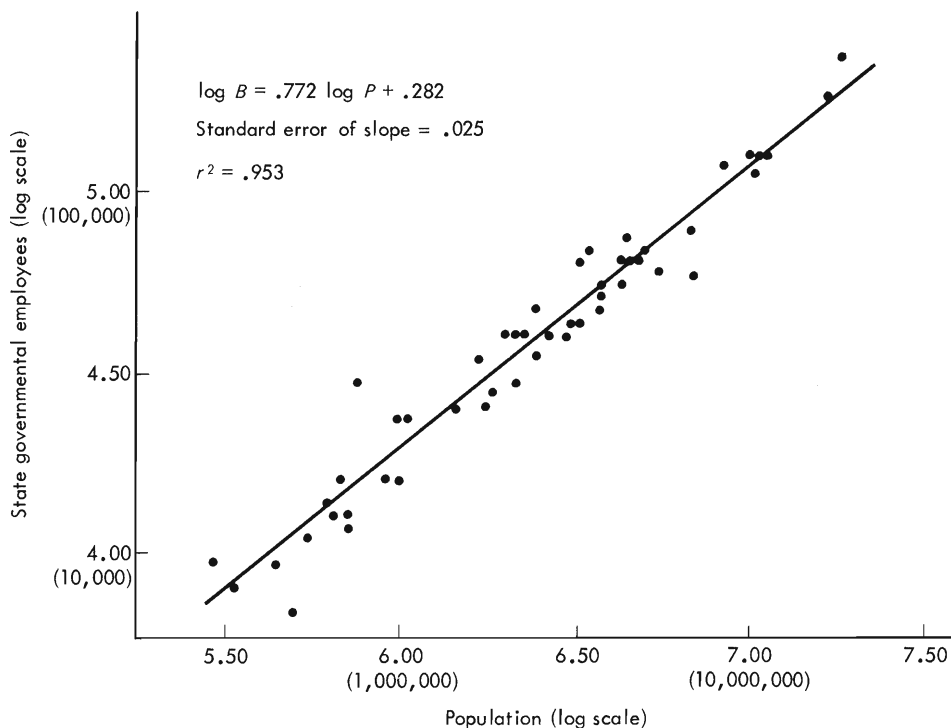


FIGURE 3-24 Population and state government employees

faster, slower, or at the same rate as the number of citizens in the state.

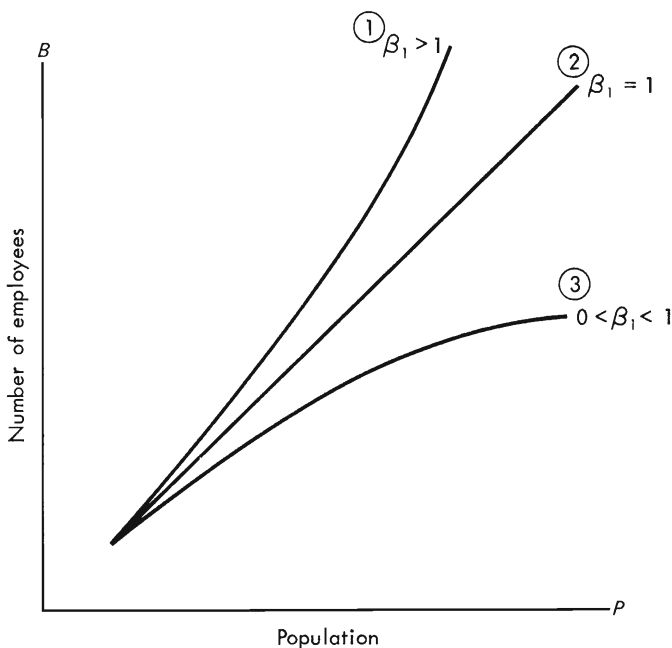
The model that helps to choose among these possibilities is

$$\log B = \beta_1 \log P + \beta_0$$

or letting $\beta_0 = \log c$ and taking antilogs puts the model in terms of the untransformed variables:

$$B = cP^{\beta_1}.$$

If β_1 is approximately one, then B approximately equals cP , which says that B grows linearly in direct proportion as P grows. In this case, there is support for what might loosely be called the "null hypothesis" concerning the relationship between size and the dependent variable. An example where β_1 would be very close to one and the null hypothesis accepted would be the relationship between the



- ① B grows faster than P
- ② B grows proportionally to P
- ③ B grows more slowly than P

FIGURE 3-25 Three types of relationships between B and P

size of the population and the number of women in the population. In this case, given the sex ratio, c would be about .52.

In terms of the untransformed variables, if the estimated regression coefficient is greater than one, the slope increases as P increases. If β_1 lies between zero and one, the slope continually decreases. Figure 3-25 shows this result in a plot of the untransformed variables.

For the fifty states, we have the following results:

$$\log B = .772 \log P + .282,$$

$$\text{Elasticity} = \hat{\beta}_1 = .772,$$

$$\text{Standard error of elasticity} = .025, \quad r^2 = .953.$$

Figure 3-24 shows the fitted curve.

The estimated elasticity is less than unity, indicating that the number of government employees grows somewhat more slowly than population. A change of one percent in the size of the population of a state is associated with a change of .772 percent in the number of government employees.

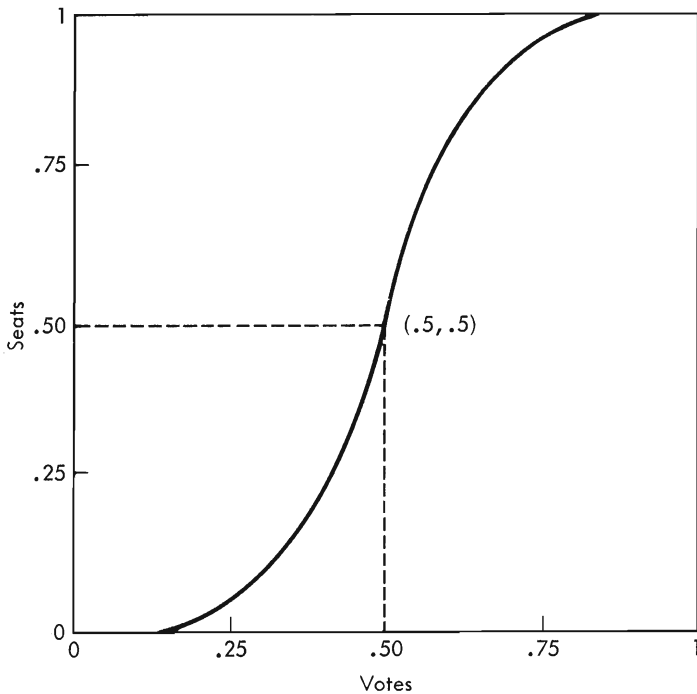
Note that the correlation coefficient is virtually useless in this problem. The square of the correlation provides a measure of the goodness of fit; but what is important is the estimate of the slope.

EXAMPLE 3 FOR THE LOG-LOG CASE: TESTING THE "CUBE LAW" RELATING SEATS AND VOTES WITH A LOGIT MODEL

One well-known description of the relationship between votes and seats in two-party systems is the "cube law."¹⁶ The most economical statement of the law is that the cube of the vote odds equals the seat odds, where the vote odds are the ratio of the share of the votes received by one party divided by the share of the votes received by the competing party. For example, if both parties win 50 percent of the votes, then the odds are one to one. Figure 3-26 shows the line traced out by the cube law.

Quite a number of papers have touched upon the law and, in the last few years, the law has enjoyed a certain vogue and has been fitted to electoral outcomes in England, the United States, New Zealand, and, in a modified form, Canada. With one or two exceptions, discussions of the law are quite sympathetic, suggesting that it is

¹⁶This discussion follows E. R. Tufté, "The Relationship Between Seats and Votes in Two-Party Systems," *American Political Science Review*, 68 (June 1973), 540-54. Additional discussion of the paper is found in the *American Political Science Review*, 68 (March, 1974), 207-13.



$$\frac{S}{1-S} = \left(\frac{V}{1-V}\right)^3$$

$$S = \frac{V^3}{1 - 3V + 3V^2}$$

S = proportion of seats for one party

$1 - S$ = proportion of seats for the other party
in a two party system

V = proportion of votes for one party

$1 - V$ = proportion of votes for other party

FIGURE 3-26 The cube law

SOURCE: Figure follows James G. March, "Party Representation as a Function of Election Results," *Public Opinion Quarterly*, 11 (Winter 1957-58), p. 524.

a useful and accurate description of electoral realities. Most studies consider no more than a few data points and conclude that the law fits rather well—although the quality of fit is usually assessed informally and no alternative fits are tried. Let us consider a direct test of the predictions of the cube law by using the log-log model. The law is

$$\frac{S}{1-S} = \left(\frac{V}{1-V}\right)^3.$$

The ratio of shares of seats and votes won by the two parties represents the odds that a party will win a seat or a vote. Taking logarithms yields

$$\log_e \frac{S}{1-S} = 3 \log_e \frac{V}{1-V},$$

and therefore in the regression of log-odds on seats against log-odds on votes,

$$\log_e \frac{S}{1-S} = \beta_0 + \beta_1 \log_e \frac{V}{1-V},$$

the cube law makes the simultaneous joint prediction that $\beta_0 = 0$ and $\beta_1 = 3$. Table 3-8 reports the results of tests of these predictions.

The table indicates that the cube law fits poorly in six of the seven trials. It fits quite well for the last eight elections in Great Britain, but otherwise its predictions are not confirmed. In short, it is not a "law." Since previous studies have not tested the exact joint predictions of the cube law (that is, $\beta_0 = 0$ and $\beta_1 = 3$) or used as extensive a collection of data, these results should be decisive in evaluating the empirical merits of the cube law.

Our previous analysis of seats and votes (Example 4) points to other defects in the cube law. The law hides important political issues because it implies that the translation of votes into seats is (1) unvarying over place and time, and (2) always "fair," in the sense that the curve traced out by the law passes through the point (50 percent votes, 50 percent seats), and the bias is zero.

As we have seen, these implications are not true. The rate of translation of votes into seats differs greatly across political systems, ranging between gains of 1.3 to 3.7 percent in seats for each 1.0 percent gain in votes. Also the results in Table 3-8 indicate that some electoral systems persistently favor a particular party; the votes-seats curve traced out by the data does not inevitably pass close by the point (50 percent votes, 50 percent seats).

The model estimated in the test of the cube law is called a "logit model." Define the odds in favor of a party winning a seat as $S/(1-S)$ and the vote odds as $V/(1-V)$. The logit model is the regression of the logarithm of seat odds against the logarithm of vote odds (a regression used earlier to test the specific predictions of the cube law):

TABLE 3-8
Testing the Predictions of the Cube Law (and Simultaneously Estimating the Logit Model)

	$\hat{\beta}_0$	$\hat{\beta}_1$	Standard error of slope	r^2	Does $\beta_0 = 0$ and $\beta_1 = 3$ as cube law predicts?	Is $\beta_0 \neq 0$; that is, is there a significant bias?
Great Britain	-.02	2.88	.30	.94	Yes	No bias
New Zealand	-.12	2.31	.27	.91	No	Yes, there is a bias
United States, 1868-1970	.09	2.52	.24	.68	No	Yes
United States, 1900-1970	.17	2.20	.15	.86	No	Yes
Michigan	-.17	2.19	.43	.76	No	Yes
New Jersey	-.77	2.09	.59	.29	No	Yes
New York	-.23	1.33	.19	.74	No	Yes

$$\log_e \frac{S}{1-S} = \beta_0 + \beta_1 \log_e \frac{V}{1-V}.$$

Since both variables are logged, the estimate of the slope, $\hat{\beta}_1$, is the estimated elasticity of seat odds with respect to vote odds; that is, a change of one percent in the vote odds is associated with a change of $\hat{\beta}_1$ percent in seat odds.

The logit model has the advantage over the linear fit used in Example 4 of producing a reasonable predicted value for the share of seats for all logically possible values of the share of votes; the predicted values stay between 0 and 100 percent seats for any percentage share of votes. As noted earlier, this is only a theoretical virtue, since the more extreme values do not occur empirically. The logit model also provides a direct test of the hypothesis that an electoral system is unbiased, since $\beta_0 = 0$ in an unbiased system. As shown in Table 3-8, there is a statistically significant bias in all cases except Great Britain.

CASE II—RESPONSE VARIABLE LOGGED, DESCRIBING VARIABLE NOT LOGGED

Here we have the model of the form

$$\log Y = \beta_0 + \beta_1 X.$$

One particularly interesting application of such a model derives from the exponential:

$$Y = ae^{bX}.$$

Taking natural logarithms and letting $c = \log_e a$ puts this model into the form of case II:

$$\log_e Y = c + bX.$$

This exponential model can be estimated by ordinary least squares, and the regression coefficient has the following interpretation:

In the model $Y = ae^{bX}$, $b \times 100$ is approximately equal to the *percent increase in Y per unit increase in X*, if b is small (say, less than .25).

The proof of this statement relies on the series expansion of e^X :

Percent increase in Y per unit increase in X

$$\begin{aligned} & \frac{\Delta Y}{Y} \\ &= \frac{\Delta Y}{\Delta X} \\ &= \frac{Y_2 - Y_1}{Y_1} \quad (\text{since } \Delta X = X_2 - X_1 = 1) \\ &= \frac{ae^{bX_2} - ae^{bX_1}}{ae^{bX_1}} \\ &= e^{(bX_2 - bX_1)} - 1 \\ &= e^b - 1 \quad (\text{since } X_2 - X_1 = 1) \\ &= [1 + b + \frac{1}{2!} b^2 + \frac{1}{3!} b^3 + \dots] - 1, \end{aligned}$$

by the expansion of e^b . So, if b is small, we can drop the higher-order terms, leaving

$$\approx (1 + b) - 1 = b.$$

Thus $b \times 100$ equals the percent increase in Y associated with a unit increase in X .¹⁷

The logarithm of the response variable is used in estimating rates of increase over time. Table 3-9 shows the gross national product of Japan from 1961 to 1970. Note the increasing absolute increase in GNP growth—GNP (the yearly absolute increase) itself increases over time. One process generating such increasing increases is a constant *percentage* growth rate—just like compound interest. What is the appropriate model for a constant percentage growth rate? Consider compound interest, at i percent per year. Beginning the first year with principal P_0 leads to principal P_t after t years:

$$P_t = P_0(1 + i)^t.$$

For example, after one year:

$$P_1 = P_0(1 + i).$$

After two years

$$\begin{aligned} P_2 &= P_1(1 + i) \\ &= P_0(1 + i)^2, \end{aligned}$$

and so on. To put this into slightly more familiar notation:

$$Y_t = Y_0(1 + i)^t.$$

Taking the logarithm of both sides

$$\begin{aligned} \log Y_t &= \log[Y_0(1 + i)^t], \\ \log Y_t &= \log Y_0 + \log(1 + i)^t, \\ \log Y_t &= \log Y_0 + t \log(1 + i). \end{aligned}$$

¹⁷An application of this interpretation is found in Philip E. Sartwell and Charles Anello, "Trends in Mortality from Thromboembolic Disorders," in Advisory Committee on Obstetrics and Gynecology, Food and Drug Administration, *Second Report on the Oral Contraceptives* (Washington, D.C.: U.S. Government Printing Office, 1969), 37-39.

Now let

$$\beta_0 = \log Y_0,$$

$$\beta_1 = \log(1 + i),$$

and we have the model

$$\log Y_t = \beta_0 + \beta_1 t$$

—that is, case II. The model is estimated by letting $Y = \log Y_t$, yielding

$$Y = \beta_0 + \beta_1 t,$$

the usual linear model.

Figure 3-27 shows the GNP of Japan plotted on both an absolute scale and a logarithmic scale. Note how, for these data, the log scale throws the data points into a straight line. The changes in the logarithm of GNP are relatively constant (Table 3-9), indicating a relatively constant percentage rate of growth over time. The line for log GNP fits considerably better than the line for absolute GNP—as the r^2 shows. The fitted line for the logarithmic case is

$$\log_{10} \text{GNP} = 1.627 + .064 t.$$

The rate of growth, i , can be estimated by going back to the original linearization of the model,

$$\beta_1 = \log(1 + i),$$

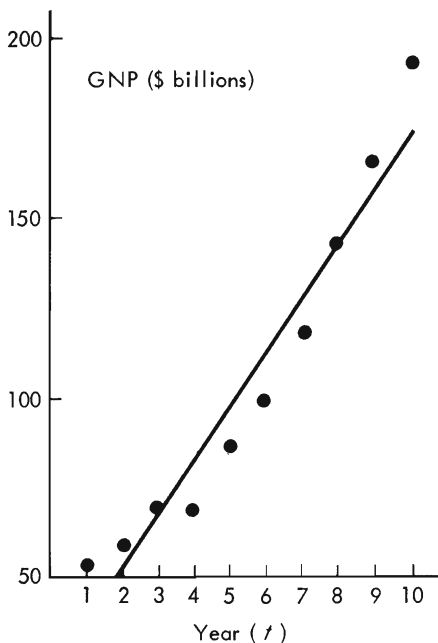
and solving by taking antilogarithms. This yields

$$\hat{i} = .159,$$

or a growth rate of almost 16 percent per year.¹⁸

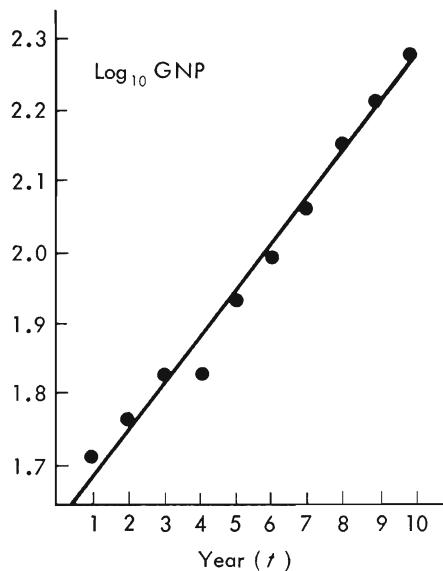
This is the yearly rate of growth. An instantaneous rate of growth can be estimated by fitting the model

¹⁸Unfortunately the estimate, \hat{i} , is biased. It does not have least-squares properties because the sum of squares was minimized with respect to log GNP rather than GNP over time.



$$\text{GNP} = 19.40 + 15.58 t$$

$$r^2 = 0.923$$



$$\text{Log}_{10} \text{GNP} = 1.627 + 0.064 t$$

$$r^2 = 0.982$$

FIGURE 3-27 Growth of GNP, Japan, 1961-1970

$$\log_e Y = \beta_0 + \beta_1 t.$$

Differentiating gives

$$\beta_1 = \frac{dY/Y}{dt},$$

the percentage rate of growth in Y .

Finally, a growth rate can be estimated quite soundly without the regression model, simply by taking the average (mean, median, or midmean) of the yearly growth rates, or the average of the logarithm.

CASE III—RESPONSE VARIABLE UNLOGGED, DESCRIBING
VARIABLE LOGGED

The model is

$$Y = \beta_0 + \beta_1 \log X.$$

TABLE 3-9
Gross National Product, Japan, 1961-1970

<i>Year</i>	<i>t</i>	<i>GNP</i> (\$ Billion)	<i>Yearly increase</i> <i>in GNP</i>	\log_{10} <i>GNP</i>	<i>Yearly increase</i> <i>in log₁₀ GNP</i>
1961	1	53		1.72	
1962	2	59	6	1.77	.05
1963	3	68	9	1.83	.06
1964	4	68	0	1.83	.00
1965	5	85	17	1.93	.10
1966	6	97	12	1.99	.06
1967	7	116	19	2.06	.07
1968	8	142	26	2.15	.09
1969	9	166	24	2.22	.07
1970	10	197	31	2.30	.08

If the logarithm of the describing variable is taken to the base 10, the regression indicates that a change in the order of magnitude of X —that is, a tenfold increase in X —is associated with a change of β_1 units in Y .

Sometimes it is useful to take the logarithm to the base 2 in this model. In such a case, the regression coefficient estimates the increase in Y when X doubles. And so when X is measured with respect to time, the estimate of the regression coefficient may be said to assess the “doubling time” of Y with respect to X . It is easy to prove that when X doubles, Y increases by β_1 units. The model is

$$Y = \beta_0 + \beta_1 \log_2 X.$$

Now suppose X doubles:

$$\begin{aligned} Y_{\text{new}} &= \beta_0 + \beta_1 \log_2 2X \\ &= \beta_0 + \beta_1 (\log_2 2 + \log_2 X) \\ &= \beta_0 + \beta_1 \log_2 X + \beta_1 \\ &= Y + \beta_1 \end{aligned}$$

—that is, the value of Y after X doubles is the old value of Y plus β_1 . Thus Y increases by β_1 units when X doubles.

Consider the following application of this model. Kelley and Mirer have developed a rule predicting how voters will vote; the predictions are made on the basis of an interview with the voter D days before the election. After the election, the voter is reinterviewed and asked how he or she voted. Thus it is possible to find the rate of error in prediction—and such errors might well be related to how many days before the election the voter was interviewed. If D were 1000 days, to take an extreme example, the error rate in prediction would be higher than if D were one day. The researchers analyzed the data first with a linear model, then with a logarithmic model:

A simple linear regression of the first of these variables on the second shows them to be strongly related. The equation yielded is:

$$\text{rate of error} = 17.4 + .23(\text{days before election}).$$

In a statistical sense this relationship explains some 28 percent of the variance in the dependent variable, and, since the standard error of the estimated coefficient is .07, the relationship is statistically significant ($t = 3.15$). Most interesting, perhaps, is the implication of the equation's constant term: Had the interviews of these respondents been conducted on election day, the mean rate of error in predicting their votes would have been 17.4 percent. . . .

And it is quite possible that this value for the constant term is too high. The volume of partisan propaganda is normally much heavier in the last two or three weeks of a presidential campaign than it is earlier. We might therefore suppose the relationship between time and changes of opinion to be like that shown in Figure [3-28], in which the likelihood of such changes (and thus the error rates of our predictions) at first increases rapidly with increases in the number of days between election day and the time the opinions were expressed, then more slowly. By regressing the rates of error in our predictions for groups of respondents on the logarithm (to the base 2) of the mean number of days before election day that the respondents in each group were interviewed, one can see if a curve like that shown in Figure [3-28] fits the data that entered into the first regression. The equation produced by this new regression is:

$$\text{rate of error} = 5.3 + 4.03(\log_2 \text{ days before election}).$$

This second equation accounts for as much of the variance in the dependent variable as did the first and yields an equally reliable estimate of the regression coefficient ($r^2 = .28$, $t = 3.14$). The value of the equation's constant term implies that our mean rate of error in predicting the votes of groups of respondents would have been 5.3 percent . . . if those respondents had been interviewed one day

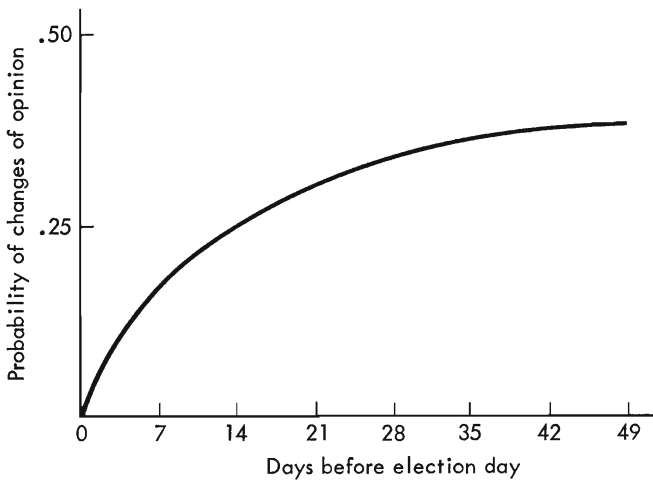


FIGURE 3-28 Hypothetical relationship between the likelihood that opinions will change and the time that attitudes toward parties and candidates are expressed

before election day. The equation as a whole implies that, starting from the day before the election, the error rate in predictions derived from the Rule will rise by four percentage points with each doubling of the length of time before election day that respondents are interviewed.¹⁹

Example 7: Regressions Aren't Enough— Looking at the Scatterplot

F. J. Anscombe has constructed a nice set of numbers illustrating why it is important to look at scatterplots along with the fitted equation.²⁰ Table 3-10 shows four sets of data. Their remarkable property is that all four yield exactly the same result when a linear model is fitted. The regression in all four cases is:

$$Y = 3.0 + .5 X,$$

$$r^2 = .667, \text{ estimated standard error of } \beta_1 = 0.118,$$

¹⁹Stanley Kelley, Jr., and Thad W. Mirer, "The Simple Act of Voting," *American Political Science Review*, 68 (June 1974), pp. 582-83.

²⁰F. J. Anscombe, "Graphs in Statistical Analysis," *American Statistician*, 27 (February 1973), 17-21. Copyright 1973 by the American Statistical Association. Reprinted by permission.

TABLE 3-10
Four Data Sets

DATA SET 1		DATA SET 2	
X	Y	X	Y
10.0	8.04	10.0	9.14
8.0	6.95	8.0	8.14
13.0	7.58	13.0	8.74
9.0	8.81	9.0	8.77
11.0	8.33	11.0	9.26
14.0	9.96	14.0	8.10
6.0	7.24	6.0	6.13
4.0	4.26	4.0	3.10
12.0	10.84	12.0	9.13
7.0	4.82	7.0	7.26
5.0	5.68	5.0	4.74

DATA SET 3		DATA SET 4	
X	Y	X	Y
10.0	7.46	8.0	6.58
8.0	6.77	8.0	5.76
13.0	12.74	8.0	7.71
9.0	7.11	8.0	8.84
11.0	7.81	8.0	8.47
14.0	8.84	8.0	7.04
6.0	6.08	8.0	5.25
4.0	5.39	19.0	12.50
12.0	8.15	8.0	5.56
7.0	6.42	8.0	7.91
5.0	5.73	8.0	6.89

SOURCE: F. J. Anscombe, *op. cit.*

mean of $X = 9.0$,

mean of $Y = 7.5$, for all four data sets.

And yet the four situations—although numerically equivalent in major respects—are substantively very different. Figure 3-29 shows how very different the four data sets actually are.

Anscombe has emphasized the importance of visual displays in statistical analysis:

Most textbooks on statistical methods, and most statistical computer programs, pay too little attention to graphs. Few of us escape being indoctrinated with these notions:

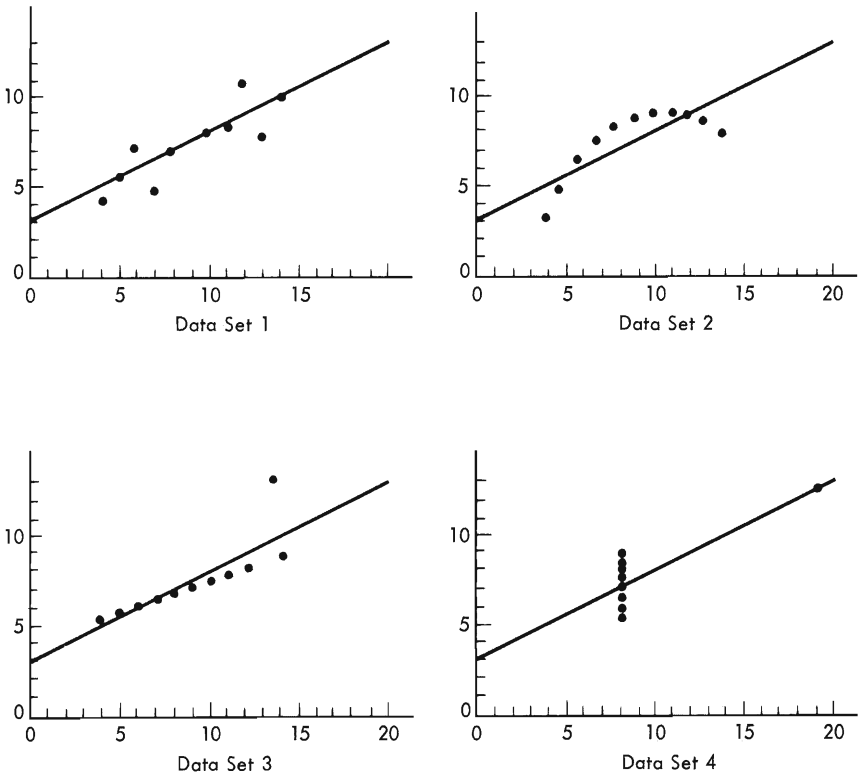


FIGURE 3-29 Scatterplots for the four data sets of Table 3-10
SOURCE: F. J. Anscombe, *op cit.*

- (1) numerical calculations are exact, but graphs are rough;
- (2) for any particular kind of statistical data there is just one set of calculations constituting a correct statistical analysis;
- (3) performing intricate calculations is virtuous, whereas actually looking at the data is cheating.

A computer should make *both* calculations *and* graphs. Both sorts of output should be studied; each will contribute to understanding.

Graphs can have various purposes, such as: (i) to help us perceive and appreciate some broad features of the data, (ii) to let us look behind those broad features and see what else is there. Most kinds of statistical calculation rest on assumptions about the behavior of the data. Those assumptions may be false, and then the calculations may be misleading. We ought always to try to check whether the assumptions are reasonably correct; and if they are wrong we ought

to be able to perceive in what ways they are wrong. Graphs are very valuable for these purposes.²¹

Up until now we have considered only one-variable explanations of the response variable. But the world is surely often more complicated than that and response variables have more than a single cause. In the next chapter, we examine the *multiple regression* model which allows us to take into account effectively several explanatory variables—at least some of the time.

²¹ Anscombe, *op. cit.*, p. 17.

Multiple Regression

“Some circumstantial evidence is very strong, as when you find a trout in the milk.”

—Henry David Thoreau

The Model

In chapter 3 we estimated the two-variable model,

$$\begin{array}{l} \text{Loss by President's} \\ \text{party in midterm} \\ \text{congressional elections} \end{array} = \beta_0 + \beta_1 \text{ (presidential} \\ \text{approval rating),}$$

and decided that a more elaborate model would help explain additional variation in the response variable. The more elaborate version used two describing variables, presidential approval and economic conditions:

$$\text{vote loss} = \beta_0 + \beta_1 \text{ (presidential} \\ \text{approval)} + \beta_2 \text{ (economic} \\ \text{conditions)}.$$

Just as in the two-variable case, we can use the data to estimate the three parameters of this model:

1. the constant term, β_0 ,
2. the regression coefficient for presidential popularity, β_1 ,
3. the regression coefficient for economic conditions, β_2 .

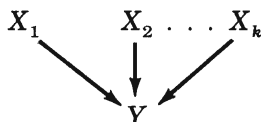
And, as before, the parameters are estimated by least squares, minimizing the sum of the squared deviations of the observed value from the fitted value:

$$\text{minimize } \Sigma (Y_i - \hat{Y}_i)^2$$

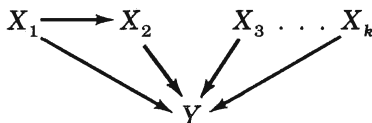
This is the multiple regression model. We can have more than two describing variables: the general multiple regression model with k describing variables is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k.$$

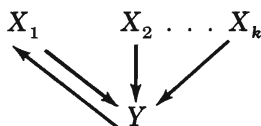
The causal model behind multiple regression is that there are k multiple, independent causes of Y , the response variable:



This is a somewhat limited model, since it excludes estimates of links *between* the describing variables—for example,



Also simple multiple regression models do not estimate feedback relationships:



Under some circumstances, models involving feedback and simultaneous relationships can be estimated.

Multiple regression is widely used in the study of economics, politics, and policy. It allows the inclusion of many describing variables in a convenient framework. It is a carefully investigated and fairly widely understood statistical procedure; thus it is a relatively effective way

to communicate the results of a multivariate analysis. And packages for running multiple regressions are available with most every computer.

Almost all of the technical apparatus used in the two-variable model applies to the multivariate case. Consider the three-variable model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2.$$

We use the data to compute:

1. the estimated regression coefficients, $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$;
2. their standard errors, $S_{\hat{\beta}_1}$, $S_{\hat{\beta}_2}$;
3. t -values to test for statistical significance of the coefficients, $\hat{\beta}_1/S_{\hat{\beta}_1}$, $\hat{\beta}_2/S_{\hat{\beta}_2}$;
4. the ratio of explained variation to total variation, R^2 .

The estimated coefficients of the model generate the predicted values

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i},$$

where X_{1i} and X_{2i} are the observed values of X_1 and X_2 , respectively, for the i th case. Now, since we have an observed and a predicted value for each observation, the residuals are defined as usual, measured along the Y axis:

$$Y_i - \hat{Y}_i,$$

and $\Sigma (Y_i - \hat{Y}_i)^2$ is minimized in the estimates of β_0 , β_1 , and β_2 . No other set of $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ will make the sum of the squared deviations smaller. As in the two-variable case, the principle of least squares generates the estimating equations for the coefficients. And, as in the two-variable case, a variety of assumptions about the data are required for the sound application of statistical significance testing in the model.¹

The percentage of the variance explained statistically is also analogous to the two-variable case:

$$R^2 = \frac{\text{explained variation}}{\text{total variation}} = \frac{\Sigma (\hat{Y}_i - \bar{Y})^2}{\Sigma (Y_i - \bar{Y})^2}.$$

¹See Ronald J. Wonnacott and Thomas H. Wonnacott, *Econometrics* (New York: Wiley, 1970); J. Johnston, *Econometric Methods*, 2d ed. (New York: McGraw-Hill, 1972); or other statistics or econometrics texts for discussion of the assumptions.

Since the R^2 provides some measure of the quality of overall fit of the describing variables in predicting Y , it is sometimes used to choose between different regressions containing different combinations of describing variables.

R can also be interpreted as the simple correlation between the observed and predicted values; that is,

$$R = r_{Y\hat{Y}}.$$

The estimated regression coefficients in a multiple regression are interpreted as *partial slopes*. They try to answer the question: When X_i , the i th describing variable, changes by one unit and all the other describing variables are held constant (in a statistical sense), how much change is expected in Y ? The answer is β_i units. If the describing variables were completely unrelated to one another, then the regression coefficients in the multiple regression would be the same as if each describing variable were regressed one at a time on Y . However, the describing variables are inevitably interrelated, and thus all the coefficients in the model are estimated and examined in combination.

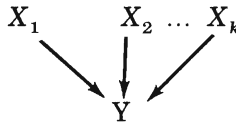
Two different types of regression coefficients—unstandardized and standardized—are used in practice. Unstandardized coefficients are interpreted in the units of measurement in which the variables are measured; for example, a one percent change in votes is associated with a β_1 percent change in seats. Standardized coefficients rescale all the variables into standard deviations from the mean:

$$\frac{Y_i - \bar{Y}}{S_Y}, \frac{X_{1i} - \bar{X}_1}{S_{X_1}}, \frac{X_{2i} - \bar{X}_2}{S_{X_2}}, \dots$$

Thus, in the standardized case, all variables are expressed in the same units—that is, in standard deviations. Standardized regression coefficients are analogous to the correlation coefficient in the two-variable case; unstandardized coefficients are analogous to the slope in the two-variable case. Standardized coefficients are useful when the natural scale of measurement does not have a particularly meaningful interpretation or when some relative comparison of the variables with respect to their standard deviations is needed. All the examples presented here use unstandardized regression coefficients.

The regression coefficients gain their meaning from the substance of the problem at hand. The statistical model merely provides the answer to the question: *Under the assumption that X_i is a cause*

of Y , what is the expected change in Y for a unit change in X_i ? Thus the estimating procedure *assumes* the causal model:



Whether or not there really is a causal relationship between Y and X_1, X_2, \dots, X_k depends on having a theory, consistent with the data, that links the variables. And in trying to assess the independent effect of one of the describing variables on Y by “holding constant” or “adjusting out” all the other describing variables, we must always keep in mind that the “holding constant” or “adjusting out” is done *statistically*, by the manipulation of the observed data. The variables are passively observed; we are not really intervening in the system and holding constant all variables except one. And so the causal structure of the multiple regression model is not strongly tested by the statistical control, adjustment, or holding constant of the variables. George Box soundly described the contrast between the statistical control of observed variables and the actual experimental control (and deliberate manipulation) of variables: “To find out what happens to a system when you interfere with it, you have to interfere with it (not just passively observe it).”²

Still, in many cases in political and policy analysis, the best we can do in trying to understand what is going on is to hold constant or control variables statistically rather than experimentally—for there is simply no other way to investigate many important questions.

Example 1: Midterm Congressional Elections— Presidential Popularity and Economic Conditions

In every midterm congressional election but one since the Civil War, the political party of the incumbent President has lost seats in the House of Representatives. This persistent outcome results from differences in turnout in midterm compared to on-year elections:

Explanation of the Administration’s loss at midterm must be sought not so much by examining the midterm election itself as by looking

²Quoted in John P. Gilbert and Frederick Mosteller, “The Urgent Need for Experimentation,” in Frederick Mosteller and Daniel P. Moynihan, eds., *On Equality of Educational Opportunity* (New York: Vintage, 1972), p. 372.

at the preceding presidential election. The stimulation of the presidential campaign brings a relatively large turnout. It attracts to the polls persons of low political interest who in large degree support the winning presidential candidate and, incidentally, his party's congressional candidates. At the following midterm congressional election, turnout drops sharply. . . . Those who stay home include in special degree the in-and-out voters who had helped the President and his congressional ticket into office. As they remain on the sidelines at midterm, the President's allies in marginal districts may find themselves voted from office. The coattail vote of the preceding presidential year that edged these Representatives into office simply stays at home . . .³

Yet this view of midterm elections is incomplete—for it only explains why the President's party should almost always be operating in the loss column rather than accounting for the *amount* of votes lost by the President's party. In statistical parlance, what has been explained is the location of the mean rather than variability about the mean. In studying the variability about the mean, we seek to answer such questions as: Why do some Presidents lose fewer congressional seats at the midterm than other Presidents? What factors affect the magnitude of the loss of congressional seats by the President's party? In Chapter 3, we used a two-variable regression to begin to answer these questions; however, that model left some variability unexplained. A more complicated model, bringing in the effect of economic conditions on the election, appears useful.

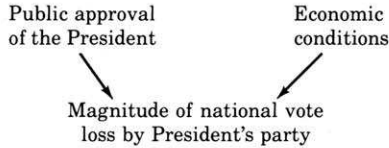
In order to explain the magnitude of the loss of votes and congressional seats by the President's party in midterm elections, we will estimate the following multiple regression model:

$$\text{Votes loss by President's party} = \beta_0 + \beta_1 \left[\begin{array}{c} \text{Presidential} \\ \text{popularity} \end{array} \right] + \beta_2 \left[\begin{array}{c} \text{Economic} \\ \text{conditions} \end{array} \right]$$

The idea is, then, that the lower the approval rating of the incumbent President and the less prosperous the economy, the greater the loss of support for the President's party in the midterm congressional elections. Thus the model assumes that voters, in midterm elections, reward or punish the political party of the President on the basis of their evaluation of (1) the performance of the President in general and (2) his management of the economy in particular.

³V. O. Key, *Politics, Parties, and Pressure Groups*, 5th ed. (New York: Thomas Y. Crowell, 1964), pp. 568–69.

The model is:



Three variables must be measured. With respect to economic conditions, recent studies of the relationship between aggregate economic conditions and the outcome of congressional elections show that interelection shifts of ordinary magnitude in unemployment have less impact on congressional elections than do shifts in real income.⁴ Thus the most meaningful measure of economic conditions for our model appears to be the interelection change in real disposable income per capita. This measure probably may reflect the economic concerns of most voters, for it assesses the short-run shift in the average economic conditions prevailing at the individual level—a shift in conditions for which some voters might hold the incumbent administration responsible.

For this model, the public's evaluation of the President's general performance is measured by the standard Gallup Poll question: "Do you approve or disapprove of the way President _____ is handling his job as President?" Table 4-1 shows responses to the survey taken each September prior to the midterm election.

TABLE 4-1
The Data

Year		Mean congressional vote for party of current President in last 8 elections	Nationwide congressional vote for party of current President	Standardized vote loss	Gallup Poll rating of President at time of election	Current yearly change in real disposable income per capita
1946	Democratic	52.57%	45.27%	7.30%	32%	-\$36
1950	Democratic	52.04%	50.04%	2.00%	43%	\$99
1954	Republican	49.79%	47.46%	2.33%	65%	-\$12
1958	Republican	49.83%	43.91%	5.92%	56%	-\$13
1962	Democratic	51.63%	52.42%	-.79%	67%	\$60
1966	Democratic	53.06%	51.33%	1.73%	48%	\$96
1970	Republican	46.66%	45.68%	.98%	56%	\$69

The most important variable to measure well is the magnitude of the vote loss by the President's party. The idea of "loss" implies the question "Relative to what?" The relevant comparison is between the normal, long-run congressional vote for the political party of the current President and the outcome of the midterm election at hand—that is, a standardized vote loss:

$$\left(\begin{array}{c} \text{standardized vote} \\ \text{loss by President's} \\ \text{party in the } i\text{th} \\ \text{midterm election} \end{array} \right) = \left(\begin{array}{c} \text{average vote for} \\ \text{party of current} \\ \text{President in the} \\ \text{last 8 elections} \end{array} \right) - \left(\begin{array}{c} \text{vote for} \\ \text{President's} \\ \text{party in the} \\ \text{}i\text{th election} \end{array} \right)$$

The loss is measured with respect to how well the party of the current

⁴Gerald H. Kramer, "Short-Term Fluctuations in U.S. Voting Behavior, 1896-1964," *American Political Science Review*, 65 (March 1971), 131-43; George J. Stigler, "General Economic Conditions and National Elections," *American Economic Review Papers and Proceedings*, 63 (May 1973), 160-67 and further discussion, 169-80.

President has normally tended to do, where the normal vote is computed by averaging that party's vote over the eight preceding congressional elections. This standardization is necessary because the Democrats have dominated postwar congressional elections; thus, if the unstandardized vote won by the President's party is used as the response (dependent) variable, the Republican presidents would appear to do poorly. For example, when the Republicans win 48 percent of the national congressional vote, it is, relatively, a substantial victory for that party and should be measured as such. The eight-election normalization takes this effect into account.

Table 4-1 shows the data matrix for the postwar midterm elections. We now consider the multiple regression fitting these data.

Table 4-2 shows the estimates of the model's coefficients. The results are statistically secure, since the coefficients are several times their standard errors. The fitted equation indicates:

1. A change in Presidential popularity of 10 percentage points in the Gallup Poll is associated with a national change of 1.3 percentage points in national midterm votes for congressional candidates of the President's party.
2. A change of \$100 in real disposable personal income per capita in the year prior to the midterm election is associated with a national change of 3.5 percentage points in midterm votes for congressional candidates of the President's party.

The fitted equation explains statistically 89.1 percent of the variance in national midterm election outcomes; or, to put it another way, the correlation between the actual election results and those predicted by the model is .944. Since the fitted equation uses two meaningful explanatory variables, it seems reasonable to believe in this case

TABLE 4-2
Multiple Regression Fitting Standardized Vote Loss by
President's Party in Midterm Elections

	<i>Regression coefficient and (standard error)</i>
β_1 Presidential approval rating (Gallup Poll, two months before election)	-.133 (.038)
β_2 Inter-election change in real disposable personal income per capita	-.035 (.015)

$$\beta_0 = 11.083, R^2 = .891.$$

that a successful statistical explanation is also a successful substantive explanation.

The multiple regression model is an equation, weighting the particular values (prevailing in a given election) of Presidential popularity and economic conditions. Thus the recipe for predicting the midterm outcome is to take .133 of the percent approving the President and .035 of the recent change in disposable personal income, subtract all this from β_0 (which is 11.083), and this gives the predicted shift in the midterm vote. Let us see how the equation worked for 1970. The equation, as shown in Table 4-2, fitted to the data is:

$$\text{standardized} = 11.083 - .133 \left(\begin{array}{c} \text{Percent approving} \\ \text{President} \end{array} \right) - .035 \left(\begin{array}{c} \text{Change in} \\ \text{income} \end{array} \right) \\ \text{vote loss}$$

For 1970, the percent approving the President was 56 percent; the change in disposable personal income per capita was \$69. Putting these particular values in the weighted combination of the regression yields:

$$\begin{aligned} \text{standardized} & \\ \text{vote loss} & = 11.083 - .133 (56) - .035 (69) \\ \text{predicted for 1970} & \\ & = 11.083 - 7.448 - 2.415 \\ & = 1.2 \end{aligned}$$

As Table 4-1 shows, the actual standardized vote loss for 1970 was 1.0, and thus the model fits the data rather well for 1970. As usual, the residual is the observed minus the predicted value; and thus the residual for 1970 from the fitted regression is -0.2 .

As another check of the adequacy of the model, its predictions of midterm outcomes were compared with those made by the Gallup Poll in the national survey conducted a week to ten days before each election. As Table 4-3 shows, the model outperforms, in six of seven elections, the pre-election predictions based on surveys directly asking voters how they intend to vote. All this, of course, is after the fact; it would be more useful to have a prediction in hand prior to the election to test the model.

An analysis based on so few data points ($N = 7$ elections) can be very sensitive to outlying values in the data. In order to test

TABLE 4-3
After-the-Fact Predictive Error of the Model

Year	Actual vote for House candidates, President's party	Gallup Poll prediction	Model prediction	Gallup absolute error	Model absolute error
1946	45.3	42	44.5	3.3	.8
1950	50.0	51	50.2	1.0	.2
1954	47.5	48.5	46.9	1.1	.6
1958	43.9	43	45.6	.9	1.7
1962	52.4	55.5	51.6	3.1	.8
1966	51.3	52.5	51.8	1.2	.5
1970	45.7	47	45.5	1.3	.2

Average absolute error, Gallup = 1.7 percentage points

Average absolute error, Model = 0.7 percentage points

the stability of the fitted equation, the multiple regression was recomputed after omitting one election at a time. Table 4-4 shows the results; even when the regression is based on six elections, the regression coefficients remain fairly stable. The greatest shift occurs when the outlying values for 1946 (very low Presidential approval ratings and a decline in real disposable income per capita in the early postwar period) are dropped from the estimation.

Does the strong aggregate responsiveness of midterm outcomes to economic conditions and evaluations of the President's performance indicate anything about the rationality of the electorate—or about, at least, that half of the eligible citizenry turns out in off-year elections?⁵ Such is the usual line of argument, for how else does one explain the choice of variables in the model and the ultimate results? It is important to realize, however, that all we are seeing in these data (and in the many similar studies) is the totally *aggregated* evidence that speaks only most indirectly to what must be the central political questions concerning the rationality of *individual* voters:

1. Do some voters make more rational calculations than others? Which voters? How many?
2. What are the components of these calculations?
3. What kinds of decision rules do individual voters use? Which voters use what decision rules?
4. What conditions encourage voter rationality?
5. How may these conditions be nurtured?

⁵ Angus Campbell, "Voters and Elections: Past and Present," *Journal of Politics*, 26 (November 1964), 745-57.

TABLE 4-4
 Re-estimating the Regression Coefficients When the Data Points are
 Omitted One at a Time

<i>Year omitted</i>	<i>Constant term</i>	<i>Presidential popularity</i>	<i>Change in economic conditions</i>	<i>R²</i>
1946	17.62	-.23	-.052	.94
1950	10.93	-.13	-.036	.89
1954	10.57	-.12	-.038	.90
1958	11.10	-.15	-.028	.99
1962	10.11	-.11	-.034	.88
1966	10.87	-.13	-.037	.89
1970	11.06	-.13	-.035	.88

Thus, although the results are impressive in terms of the large R^2 , there are still substantial inferential problems in trying to interpret the meaning of the model—since the data do not speak directly to the explanatory mechanism postulated to explain the findings.

Let us consider the steps in the construction of this regression in order to look at some of the broader issues in constructing explanatory models. The steps were these:

1. A model, based on prior research and some general ideas, was specified. The model included two basic variables, presidential popularity and economic conditions. There were also several other variables that were candidates for inclusion in the model: whether the nation was involved in a war at the time of the election, the magnitude of the victory of the President's party in the preceding election, and a few others.
2. Each variable in the model was operationalized; that is, a numerical measure for the concept was found. The construction of appropriate measures required some further thought, especially with respect to the response variable, the standardized vote.
3. Several economic variables were included in the initial analysis—changes in unemployment, inflation, GNP per capita, and real disposable personal income per capita. From the beginning, the change in real disposable personal income per capita made the most substantive sense, and it turned out that it led to the most successful explanatory model in terms of variance explained. A variety of different regressions were computed.

There is, then, an interplay between explanatory ideas and the examination of the data. Some variables were tried out on the basis of a vague idea and were then discarded when they yielded no explanatory return. For example, some regressions included a variable indicating whether the nation was involved in a war (Korea or Vietnam)

during the midterm congressional election—on the hypothesis that there might be a “rally round the flag” effect helping the President. Such appeared to be the case—and the sign of the regression coefficient was in the expected direction—but the results just did not seem solid enough to warrant inclusion in the final model, especially since there are only seven data points and also since only two explanatory variables do so well.

Now, looking at several different multiple regressions and sorting around through different variables may not fit some abstract models of scientific research procedure—but it is normally done in constructing explanatory models, and it is precisely this sorting through of various notions that is the heart of data analysis. The final model reported here has gained inferential strength as a consequence of this directed search through a variety of ideas because the model has been tested against many other alternative possibilities and has survived. The strength of such an interplay between theory and data has been strongly put by Jacob Viner:

If there is agreement that relevance is of supreme importance for economic theory, it leads to certain rules of guidance as to the procedure we should follow in constructing our theoretical models. It is common practice to start with the simplest and the most rigorous model, and to leave it to a later stage, or to others, to introduce into the model additional variables or other complicating elements. I venture to suggest that the most useful type of “first approximation” would often be of a radically different character. It would consist of a listing of all the variables known or believed to be or suspected of being of substantial significance, and corresponding listing of types and directions of interrelationship between these variables. A second stage of analysis would consist of a combing out on the basis of such empirical evidence as can be accumulated of the probably least significant variables and interrelationships between variables. Instead of beginning with rigor and elegance, only from this second stage on would these become legitimate goals, and even then for a time they should be distant goals, to be given high value only after it is clear that they can be reached without substantial loss of relevance.

Such procedure, it would seem to me, would have some distinct advantages as compared to the more usual procedure on the part of theorists of starting—and often ending—with models that gain their rigor at the cost of unrealistic simplification. In the first case, important variables would be less likely to be omitted from consideration because of oversight, traditional practice, difficulty of manipulation, or unsuitability for specific types of analytical manipulation to which the researcher has an irrational attachment. Secondly, there would be at least partial awareness of what variables had been omitted from the final analysis, and therefore greater likelihood than at present that the conclusions will be offered with the qualifications

and the caution that such omission makes appropriate. Third, if the presentation of the final results includes a statement with respect to the omitted variables and the reasons for their omission, the reader of such presentation is in better position to appraise the significance of the findings and is afforded some measure of guidance as to the further information and the new or improved techniques of analysis that would be most helpful.

The final outcome of such a change in analytical procedure might well be a definite loss in rigor and elegance at least for a long time, on the one hand, but a definite gain in scope for the useful exploitation of new information and of wisdom and insight on the other hand. Such a result, I hope and believe, would in most cases constitute a new gain in relevance for understanding of reality and for the promotion of economic welfare by means of economic theorizing.⁶

Example 2: Equality of Educational Opportunity and Multicollinearity

Modern statisticians are familiar with the notions that any finite body of data contains only a limited amount of information, on any point under examination: that this limit is set by the nature of the data themselves, and cannot be increased by any amount of ingenuity expended in their statistical examination: that the statistician's task, in fact, is limited to the extraction of the whole of the available information on any particular issue.⁷

—R. A. Fisher

If two or more describing variables in an analysis are highly intercorrelated, it will be difficult and perhaps impossible to assess accurately their independent impacts on the response variable. As the association between two or more describing variables grows stronger, it becomes more and more difficult to tell one variable from the other. This problem, called "multicollinearity" in the statistical jargon, sometimes causes difficulties in the analysis of nonexperimental data.

For example, if, in Chapter 1, density and inspections (the two describing variables for the response variable of traffic fatalities) were highly associated—say, all states above a certain density had inspections and all below did not—then it would be very difficult to discover if inspections made a difference because the effect of inspections would be confounded with the effect of density. The

⁶Jacob Viner, "International Trade Theory and Its Present Day Relevance," in *Brookings Lectures, 1954, Economics and the Public Policy*. © 1955 by the Brookings Institution, Washington, D.C., pp. 128–30.

⁷R. A. Fisher, *The Design of Experiments*, 8th ed. (London: Oliver and Boyd, 1966), p. 40.

scatterplot, in this hypothetical example, would resemble Figure 4-1. In such a case there is insufficient independent variation in the two describing variables; in particular, there is a shortage of thickly populated states without inspections and thinly populated states with inspections. Without such conditions prevailing in at least a few states, the independent effect of inspections and the independent effect of density on the death rate could not be assessed.

Sometimes clusters of variables tend to vary together in the normal course of events, thereby rendering it difficult to discover the magnitude of the independent effects of the different variables in the cluster. And yet it may be most desirable, from a practical as well as scientific point of view, to disentangle correlated describing variables in order to discover more effective policies to improve conditions. Many economic indicators tend to move together in response to underlying economic and political events. Or consider a research design seeking to assess the effects of air pollution on the health of a city's residents. Such a study might be based on three areas in a city—one with badly polluted air, one with moderate pollution, and (if it could be found) one with relatively clean air. But chances are that the poor are more likely to find housing only in those unpleasant parts of

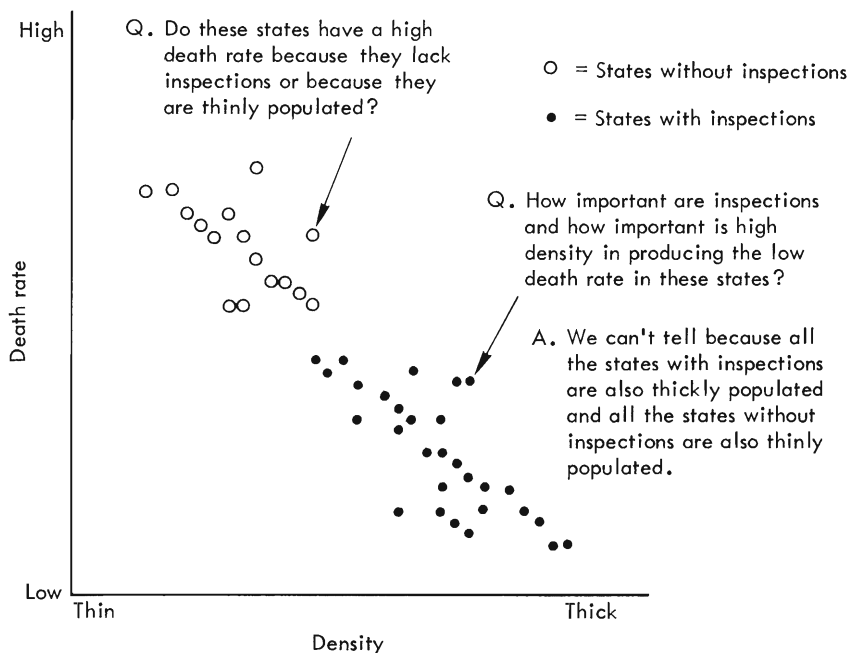


FIGURE 4-1 Hypothetical data showing collinearity between density and inspections

the city near factories and highways producing very polluted air; the moderately polluted area is more likely to be the home of those with moderate incomes; and the wealthy will be concentrated in areas relatively free of pollution. In such a situation, then, the effects of air pollution on health are confounded with the effects of income and housing on health.

The problem of multicollinearity involves a lack of data, a lack of information. In the first example, there were no *thinly* populated states *with* inspections (and vice versa); in the study of the health effects of air pollution, we lacked information about rich neighborhoods with polluted air and poor areas with fresh air.

Recognition of multicollinearity as a lack of information has two important consequences:

1. In order to alleviate the problem, it is necessary to collect more data—especially on the rarer combinations of the describing variables.
2. No statistical technique can go very far to remedy the problem because the fault lies basically with the data rather than the method of analysis. Multicollinearity weakens inferences based on *any* statistical method—regression, path analysis, causal modeling, or cross-tabulations (where the difficulty shows up as a lack of deviant cases and as near-empty cells).

Figure 4-2 shows how, when two describing variables are highly intercorrelated, a control for one variable reduces the range of variation in the other.

Since multicollinearity affects our ability to assess the independent influence of each describing variable, its consequences in the multiple regression model include increased errors in the estimate of the regression coefficients. The variance of the estimate of the regression coefficient, $\hat{\beta}_i$, is given by:

$$\text{variance of } \hat{\beta}_i = \frac{1}{N - n - 1} \frac{S_Y^2}{S_{X_i}^2} \frac{1 - R_Y^2}{1 - R_{X_i}^2},$$

where N = number of observations,

n = number of describing variables,

S_Y^2 = variance of Y ,

$S_{X_i}^2$ = variance of X_i ,

R_Y^2 = squared multiple correlation for the regression

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n,$$

$R_{X_i}^2$ = squared multiple correlation for the regression

$$X_i = \beta'_0 + \beta'_i X_i + \dots + \beta'_{i-1} X_{i-1} + \beta'_{i+1} X_{i+1} + \dots + \beta'_n X_n$$

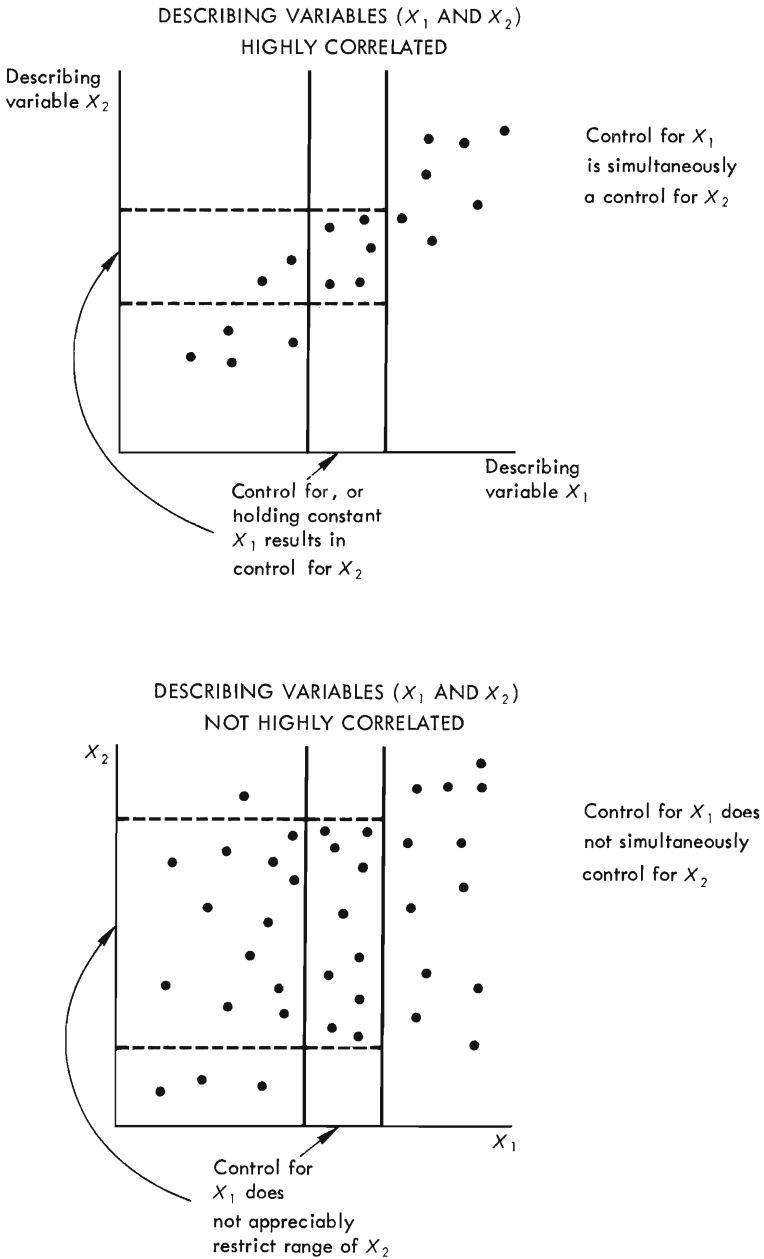


FIGURE 4-2 Effect of controlling for a variable when describing variables are strongly correlated

—that is, the regression of the *i*th *describing* variable on all the other *describing* variables.

This equation repays study. The key element is:

the variance of $\hat{\beta}_i$ is proportional to $\frac{1}{1 - R_{X_i}^2}$.

Now $R_{X_i}^2$ is the R^2 for the regression of the *i*th describing variable on all the other remaining describing variables—that is, $R_{X_i}^2$ assesses how well the describing variable X_i is explained by the *other* describing variables. So if X_i is strongly entangled with one or more of the other describing variables, $R_{X_i}^2$ will be large, close to 1.0. Consequently $1/(1 - R_{X_i}^2)$ and the variance of $\hat{\beta}_i$ will grow larger as $R_{X_i}^2$ approaches 1.0. And so the estimate of $\hat{\beta}_i$ grows more insecure as $R_{X_i}^2$ approaches closer to 1.0.

Although multicollinearity is sometimes viewed as a problem of the intercorrelation of two describing variables, it can be seen here that the variances of the estimated regression coefficients will be big *whenever* $R_{X_i}^2$ is large—which can result from a high intercorrelation between two of the describing variables *or* from a combination of three or more of the describing variables accurately predicting another describing variable. Note the variance of $\hat{\beta}_i$ is infinite when $R_{X_i}^2$ is unity (that is, when a describing variable X_i is perfectly predicted by one or more of the other describing variables). In this case, of course, it is literally impossible to tell X_i from another describing variable or combination of other describing variables. The equation for the variance of $\hat{\beta}_i$ also shows that the variance of the estimates of the regression coefficients will decrease as additional data are collected (as N grows larger).

In summary, the symptoms of multicollinearity in regression analysis include:

1. high intercorrelations between the describing variables,
2. large variances in the estimates of the regression coefficients,
3. large $R_{X_i}^2$,
4. large $R_{X_i}^2$ coupled with statistically nonsignificant regression coefficients,
5. large changes in the values of estimated regression coefficients when new variables are added to the regression, and
6. inability of computer program to compute regression coefficients (which occurs only in very severe cases of multicollinearity—in most cases the estimation procedures produce numbers as usual).

Cures for multicollinearity can sometimes be found in the following list:

1. Collect additional data, concentrating on gathering information that will alleviate the difficulty. In some research contexts, this may involve seeking information on deviant cases or special combinations of the describing variables. Johnston cites an econometric example: "Early demand studies, for example, which were based on time-series data, often ran into difficulties because of the correlation between the explanatory variables, income and prices, plus the often inadequate variation in the income series. The use of cross-section budget data, however, gives a wide range of income variation, thus permitting a fairly precise determination of the income coefficient, which can then be employed in the time-series analysis."⁸
2. Give up on nonexperimental data and consider research designs in which the key variables can be systematically varied or at least randomized out. Do experiments.
3. Remove some of the variables from the regression that are causing the trouble. For example, if two of the describing variables are highly correlated, compute regressions with only one of the variables present at a time. Or combine the variables into a summary measure (less often an approved strategy). These steps should be taken only if they make good substantive sense.

Although the use of additional information and special statistical techniques may at times alleviate the problem, it often happens in social research based on "experiments" performed by nature that it will be difficult to obtain the independent variation necessary to assess the independent effects of the describing variables. Thus some theories that assert the importance of one variable over another, while theoretically testable, are actually incapable of being tested in the face of multicollinearity.

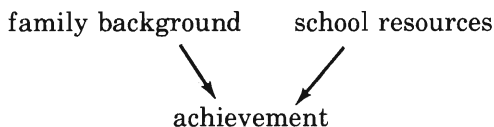
Finally, it is important to be clear about the signs of multicollinearity and just when it is a genuine threat to the validity of a study. It is not a sound or a fair statistical criticism to cry "multicollinearity" to discredit every analysis involving three or more variables.

A multicollinearity problem arose in the report on *Equality of Educational Opportunity* by James Coleman and others.⁹ The model used seeks to explain student achievement in school (as measured

⁸J. Johnston, *Econometric Methods*, 2d ed. (New York: McGraw-Hill, 1972), p. 164.

⁹James Coleman, Ernest Q. Campbell, Carol J. Hobson, James McPartland, Alexander M. Mood, Frederic Weinfeld, and Robert L. York, *Equality of Educational Opportunity* (Washington, D.C.: Office of Education, 1966). Parts of the report are reprinted in E. R. Tufte, ed., *The Quantitative Analysis of Social Problems* (Reading,

by test scores) with two clusters of variables, measures of family background aiding children in their schoolwork (such as books in the home) and measures of school resources such as the teacher-student ratio and the number of books per student in the library. In compressed form, the model is:



The analysis proceeded by first regressing achievement against the family background variables, which yielded an R^2 . Then a new regression was computed that included the school resources variables as well as family background, yielding a coefficient of R'^2 . The difference,

$$R'^2 - R^2,$$

was taken as measure of the effect of school resources on educational achievement. Although using the increase in the percent of variance explained as a measure of school resource effects on education did not ultimately compromise the main findings of the study, the method would tend to underestimate school effects somewhat and received criticism. Bowles and Levin wrote:

The most severe deficiency of the regression analysis is produced by the addition to the proportion of variance in achievement scores explained (addition to R^2) by each variable entered in the relationship as a measure of the *unique* importance of that variable. For example, assume that we seek to estimate the relationship between achievement level, Q , and two explanatory variables, X_1 and X_2 . The approach adopted in the Report is to first determine the amount of variance in Q that can be statistically explained by one variable, say X_1 , and then to determine the amount of variation in Q that can be explained by both X_1 and X_2 . The increment in explained variance (i.e., the change in the coefficient of determination, R^2) associated with the addition of X_2 to the explanatory equation is the measure used in the Report for the unique effect of that variable on Q . Thus, if X_1 explained 30 percent of the variance in Q and X_1 and X_2 together explained 40 percent, the difference, or 10 percent, is the measure of the unique effect of X_2 .

If X_1 and X_2 are completely independent of each other (orthogonal), the use of addition to the proportion of variance explained as a measure of the unique explanatory value of X_1 and X_2 is not objectionable. X_1 will yield the same increment to explained variance whether it is entered into the relationship first or second, and vice versa. But when the explanatory variables X_1 and X_2 are highly correlated with each other, as are the background characteristics of students and the characteristics of the schools that they attend, the addition to the proportion of variance in achievement that each will explain is dependent on the order in which each is entered into the regression equation. By being related to each other, X_1 and X_2 share a certain amount of explanatory power which is common to both of them. The shared portion of variance in achievement which could be accounted for by either X_1 or X_2 will always be attributed to that variable which is entered into the regression first. Accordingly, the explanatory value of the first variable will be overstated and that of the second variable understated.

The relevance of this problem to the analysis in the Report is readily apparent. The family background characteristics of a set of students determine not only the advantages with which they come to school; they also are associated closely with the amount and quality of resources which are invested in the schools. As a result, higher status children have two distinct advantages over lower status ones: First, the combination of material advantages and strong educational interests provided by their parents stimulate high achievement and education motivation; and second, their parents' relatively high incomes and interest in education leads to stronger financial support for and greater participation in the schools that their children attend. This reinforcing effect of family background on student achievement, both directly through the child and indirectly through the school, leads to a high statistical correlation between family background and school resources.

The two sets of explanatory variables are so highly correlated that after including one set in a regression on achievement, the addition to the fraction of total variance explained (R^2) by the second set will seriously understate the strength of the relationship between the second variables and achievement. Yet the survey made the arbitrary choice of first "controlling" for student background and then introducing school resources into the analysis. Because the student background variables—even though crudely measured—served to some extent as statistical proxies for school resources, the later introduction of the school resource variables themselves had a small explanatory effect. The explanatory power shared jointly by school resources and social background was thus associated entirely with social background. Accordingly, the importance of background factors in accounting for differences in achievement is systematically inflated and the role of school resources is consistently underestimated.¹⁰

¹⁰Samuel Bowles and Harry Levin, "The Determinants of Scholastic Achievement—An Appraisal of Some Recent Evidence," *Journal of Human Resources*, 3 (© 1968 by the Regents of the University of Wisconsin), pp. 14–16.

Example 3: A Five-Variable Regression— The Size of Democratic Parliaments

Here we examine a five-variable multiple regression that illustrates the following statistical points:

- taking logarithms to test a “cube root law” by converting the law into a linear model,
- interpreting a regression coefficient as an elasticity,
- using R_i^2 as a check for multicollinearity,
- using a “dummy variable” so that a dichotomous, categoric variable can be included in a regression,
- interpreting R^2 as the square of the correlation between the observed and predicted values of the response variable.

The multiple regression reported here evaluates some of factors determining parliamentary size—the number of representatives in the lower house—in twenty-nine relatively democratic countries of the world. Parliaments differ greatly in size; Liechtenstein’s Diet has 15 deputies, the Italian Chamber of Deputies has 630 members, West Germany’s Bundestag 496, the French National Assembly 481, and Sweden’s new unicameral parliament 350. Some large countries have relatively small parliments: India, with a population $2\frac{1}{2}$ times that of the United States, has 500 deputies sitting in its House of the People with each deputy representing, on average, over one million citizens. At the other extreme, the 19,000 residents of San Marino have a 60-member Great and General Council—resulting in one representative for every 320 citizens.

The number of representatives elected to parliament determines, in part, the extent to which local interests are represented at the national level; larger parliaments, other things (especially population) being equal, permit a more precise representation. However, in larger parliaments each member not only has an arithmetically smaller voice, but also larger parliaments typically have greater centralization of leadership and more rules limiting the conduct of their members both in debate and in the diversity of their concerns. These two conflicting factors—the representation of citizens and the manageability of the chamber—must be resolved by the framers of new constitutions. Many constitutions of the eighteenth and nineteenth centuries specified a particular ratio of citizens to representatives, and parliamentary size grew right along with the population.

As a consequence of this mixture of factors, parliamentary size is closely linked to population: the more populous countries have larger parliaments. Dodd proposed a "cube-root law":

$$\text{number of members of parliament} = (\text{population})^{1/3}.$$

Dodd correlated these variables and found that the cube root of population explained 67 percent of the variation in parliamentary size for 55 nations in 1950.

Dodd's model may be written by taking logarithms

$$\log \text{ members} = 1/3 (\log \text{ population}).$$

Thus in the regression of members (log) against population (log),

$$\log \text{ members} = \beta_1 (\log \text{ population}) + \beta_0,$$

the cube-root law predicts that $\beta_1 = 1/3$ and $\beta_0 = 0$. Confirming these predictions provides a better test of the law than merely correlating the cube root with the size of parliament. That correlation does not test the specific hypothesis of the law; it merely supports the general proposition that there is a relationship between the two variables. Taking the logarithms of both variables also, as we saw in Chapter 3, yields a useful interpretation of the slope of the fitted line. The estimate of the slope, β_1 , measures the *percentage change* in the size of parliament associated with a *change of one percent* in the size of the population: β_1 is the least-squares estimate of the elasticity of parliamentary size with respect to population.

Here Dodd's law is tested with data from twenty-nine of the more democratic countries in the world in 1970. The multiple regression in Table 4-5 shows that the population elasticity of parliamentary size is .44, indicating that if a county was one percent above the average in population size, it was typically .44 percent above the average in parliamentary size. The standard error of the estimated elasticity is quite small, .022; thus the estimate of the elasticity itself, .44, quite surely differs from the prediction of the cube-root law.

There are three other describing variables in the regression shown in Table 4-5:

Population growth rate Although many of the democracies have relatively low growth rates, there is still sufficient variability to explain differences in parliamentary size. Countries that are growing rapidly in population size tend to have smaller parliaments, other things being equal, than countries growing more slowly. When all the other describing variables in the equation are fixed at their means, a change of one percentage point in growth rate from an annual rate of one percent to two percent across countries is associated with a decrease in the size of parliament from 196 seats to 144 seats.

TABLE 4-5
Parliamentary Size (logarithm) for Twenty-nine Democracies

	<i>Regression coefficient</i>	<i>Standard error</i>	R^2_i
Population (log)	.440	.022	.14
Annual population growth rate	-.135	.020	.13
Number of political parties	.051	.013	.26
Bicameral—unicameral	.066	.040	.20
$R^2 = .952$			

All coefficients are statistically significant at the .001 level, with the exception of the variable bicameral—unicameral. That coefficient is significantly different from zero at the .06 level.

Number of parties in the party system The greater the number of parties in the present-day party system, the larger the parliament. Other things being equal, two-party systems have parliaments averaging about 137 seats; multiparty systems, 195 seats. A larger party system may reflect somewhat greater underlying diversity in the society, and the constitutional framers may then create a larger than normal parliament in an effort to represent that diversity. Perhaps a more plausible explanation is that in a multiparty system, many parties will participate in the bargaining over parliamentary size and the smaller parties will work hard for a large-sized parliament, so that at least some of their party officials will be able to hold parliamentary seats. Parliaments sufficiently large to include the leading officials of each party may be quite inflated in size, particularly if the votes of the minor parties are scattered. If such a process operated for a number of years as the distribution of seats shifted from party to party, then the incumbent parliamentarians might well favor increases in the size of parliament so that they or their colleagues would stay in office even with some shifts in the share of votes received by each party.

Bicameral—unicameral parliaments Unicameral parliaments are typically somewhat larger than the lower chambers of bicameral parliaments. Some unicameral systems have come about from a merger of two chambers; here the interests of incumbent parliamentarians are obvious. Other things being equal, the unicameral parliaments average 189 seats in the fitted model; the lower chamber of bicameral parliaments, 163 seats.

The numerical coding for this variable was:

bicameral = 0,
unicameral = 1.

Such a dichotomous categoric variable is called a "dummy variable," and such variables are used to include categoric variables in multiple regression models. The following are examples of dummy variables:

REGION

0 = North

1 = South

CHANGE IN A TIME SERIES

0 = before tax cut,

1 = after tax cut

SEX

0 = male,

1 = female

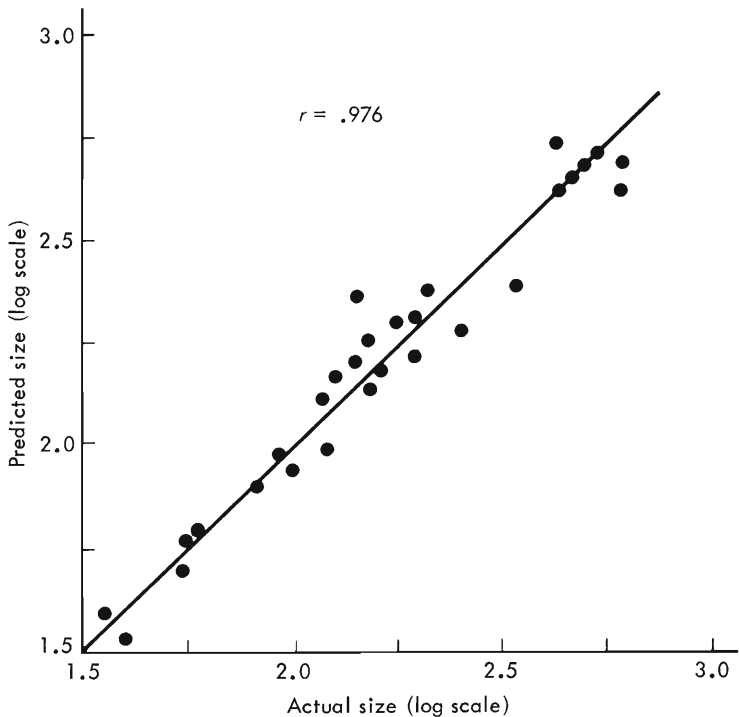


FIGURE 4-3 Actual and predicted parliamentary size, twenty-nine democracies

Table 4-5 shows the value of R_i^2 , the value of R^2 resulting from the regression of the i th describing variable on all the other describing variables. The values are quite small, indicating that multicollinearity is not a problem here.

Figure 4-3 shows the relationship between the observed and predicted values of the response variable, the logarithm of parliamentary size. The correlation, $r_{Y\hat{Y}}$, between the observed and predicted values is .976. That value squared is mathematically equivalent to R^2 , the proportion of variance in the logarithm of parliamentary size explained by the regression:

$$r_{Y\hat{Y}}^2 = (.976)^2 = .952 = R^2.$$

How was the regression reported in Table 4-5 chosen? At the start of the analysis, six describing variables were considered as possible candidates for inclusion in the final model. In addition to the four variables already discussed, two others were considered good candidates: whether or not the country was in Europe and the institutional age of the currently established parliament. It appeared that European countries, for one reason or another, had large parliaments. The length of time the parliament had been established under the current constitution was included as a possible describing variable on the speculation that older parliaments might be larger. Table 4-6 shows twelve different multiple regressions using various combinations of the six candidate describing variables. Let us look through these twelve different regressions to see the search for the model previously reported in Table 4-5. It will be clear that several different models could have been the model of choice.

TABLE 4-6
Twelve Regressions Explaining Parliamentary Size (Log)

Describing variables	Regression number											
	1	2	3	4	5	6	7	8	9	10	11	12
Population size (log)	.41	.40	.42	.41	.38	.38	.40	.39	.40	.41	.43	.44
Population growth rate		-.16		-.10		-.13	-.13	-.06		-.07	-.13	-.14
Bicameral—unicameral							.12		.11	.12		.07
Number political parties											.06	.05
European—not European			.26	.13	.20			.13	.20	.13		
Age of current parliament					.17	.13	.11	.12	.18	.13		
R^2	.760	.900	.891	.912	.916	.908	.928	.923	.934	.941	.946	.952
Number of describing variables	1	2	2	3	3	3	4	4	4	5	3	4

The numbers shown in the table are regression coefficients for each regression. Each of the twelve columns shows a different regression.

Regression 1 is simply the two-variable regression of parliamentary size (log) against population size (log). The regression coefficient reported in Table 4-6 indicates that a change of one percent in population was associated with a change of .41 percent in parliamentary size; 76 percent of the variance was statistically explained. Regressions 2 and 3, both with two describing variables, send the R^2 up to about 90 percent. Either the population growth rate or the country's geographic location in or out of Europe adds an additional 14 percent to the variance explained in the first regression. This suggests that we can go much farther with a model that includes both the location and the growth rate along with population size. This is regression 4; and it doesn't work. Little additional variance is picked up—and also there is a multicollinearity problem. Note how the regression coefficients on growth and European location have shifted from their previous values in regressions 2 and 3, respectively.

This is a sign of multicollinearity, confirmed by the correlation of $-.77$ (European countries have low population growth rates) between the two variables.

Regressions 5 through 9 try out various combinations of the describing variables. These trials verify the multicollinearity problem with respect to the European location variable and raise some doubts about the effectiveness of the age variable. Throwing in every variable examined so far gives regression 10, which picks up 94.1 percent of the statistical variation in parliamentary size (log) but with some problems. The European location variable is quite bothersome by now, in part because of multicollinearity but also—and more importantly—what does it mean, anyway? It is vague; such a regional variable doesn't tell us much substantively. What is it, *specifically*, about location in Europe that makes for big parliaments? So, regression 10 is about the best that can be done with the current candidate variables.

The last two multiple regressions try out a new candidate variable, the number of political parties in a country. Regression 11 reports a simple model with only three describing variables that outperforms—at least in terms of R^2 —all the previous models, including those that contain more variables. It is a parsimonious model and a relatively successful one in terms of R^2 . Regression 12 adds one more variable—the dummy variable on whether the parliament is unicameral or bicameral—to take the variance explained up to 95.2 percent.

What we have seen here is an empirical search through a variety of theoretically plausible models. The search started with some candidate variables, which were suggested by our political and historical understanding of what factors might affect this particular characteristic—size—of a political institution. The search was conducted with a variety of criteria for evaluating the different models that turned up: certain substantive criteria (for example, in part, the grounds for rejection of the European location variable) and certain statistical criteria (the statistical significance of individual regression coefficients, the value of R^2 , and multicollinearity). Now these criteria are not “merely” statistical matters, for the statistical criteria used in the choice of the models inform us about the quantitative *quality* of the model under examination. Or, more precisely, the statistical criteria help evaluate the quantitative quality of different models within the theoretical and substantive context of the search for models. The context is vital; the best statistical techniques can't rescue theoretical models that are poor, unintelligent, or misguided.

Table 4-6 also shows one of the sad facts of building complex

explanations of most political, economic, and social phenomena: often a variety of models will fit the same data relatively well. That is, the empirical evidence that is available does not always allow one to choose among different models that seek to explain the response variable. In this case, regressions 11 and 12 both do rather well; but even regressions 2 and 3 seem relatively acceptable. It is probably fair to say, however, that regressions 11 and 12 are pretty much the best among the lot. Both regressions are quite effective in predicting—and explaining—parliamentary size (log) as Table 4-5 indicated.

Table 4-6 does not, fortunately, show all possible combinations of describing variables. With six describing variables, there are a grand total of 63 different regressions involving combinations of one or more describing variables. In general, with K describing variables, there are $2^K - 1$ possible regressions. Some regression programs can, in fact, search through all possible combinations to find one or more "best" regressions. Although such searches may seem rather like brute-force empiricism (and they often are!), the criteria of choice for the best regression or regressions are intelligent and may provide a reasonable guide—when combined with substantive understanding—in searching for models. Some elegant computer programming has enabled one regression program to examine quickly every regression in cases with up to 12 describing variables—that is, 4,095 regressions.¹¹ The view is: If you're going to search for a model, why not search thoroughly?

Of course, we would trade all those searches in for one good idea. And that idea might come from looking at the data.

¹¹Cuthbert Daniel and Fred Wood, *Fitting Equations to Data* (New York: Wiley, 1971).

Index

- Accident Facts* (National Safety Council), 8 *n*
- Accident proneness, automobile, 60–62
- Accidents, automobile (*See also* Deaths—traffic):
- of high vs. low mileage drivers, 61–62
 - proneness to, prediction of, 60–62
 - safety inspections and, 5
- Acton, Forman S., 108 *n*
- Adjusted significance test, 48
- Adjustment method:
- in causal explanation, 24–28
 - description and, 2
- African tribes, 87
- Age distribution, lung-cancer deaths and, 82–83
- Air pollution, 82
- Alabama, 23
- Alker, Hayward, 112 *n*
- Almanac of American Politics, The*, 169
- American Academy of Ophthalmology and Otolaryngology, 4
- American Almanac, The*, 167
- American Association for the Advancement of Science, 166
- Amoia, Richard P., 167
- Analysis:
- cross-section, in auto-safety inspection analysis, 7–29
 - residual, 81–84
 - time-series, 6–7, 153
- Anello, Charles, 126 *n*
- Anscombe, F. J., 103 *n*, 132–33
- Arizona, 8, 23
- Arkansas, 23
- Asbestos, 84
- Association, 3
- Atomic radiation, study on effects of, 3–4
- Australia, 82, 86
- Automobile accidents:
- fatal, *see* Deaths—traffic
 - proneness to, prediction of, 60–62
 - safety inspections and, 5
- Automobile insurance companies, 60
- Automobile safety inspections, 5–29
- Automobile safety programs, 7
- Averages, weighted, 45–46

- Bantu, 87
- Barometric electoral districts (bellwether districts), 46-55
 predictive performance of (1936-1964), 48-52
- Barone, Michael, 169
- Bayes' theorem, 38
- Bean, Louis, 47
- Bell, Daniel, 31 *n*
- Bellwether electoral districts, 46-55
 predictive performance of (1936-1964), 48-52
- Bias in legislative seats-votes relationship in two-party systems, 94-97
- Bowles, Samuel, 154-55
- Box, George, 139
- Bureaucracy, population size and size of governmental, two variables logged, 119-21
- Burnham, Walter Dean, 167, 168
- Bush, Robert R., 34-35
- California, 23
 automobile-driver record study in, 61-62
- Campbell, Angus, 75 *n*, 145 *n*
- Campbell, Donald T., 6, 7 *n*
- Campbell, Ernest Q., 153 *n*
- Canada, 60, 82, 86
- Cancer:
 smoking and lung, 78-88
 test for, 37-40
- Causal explanation, 2-5
 in automobile safety inspections study, 5-29
- Causality, regression analysis and, 69, 72
- Census Bureau, 166
- Chester, Lewis, 40 *n*
- Chincha Islands, 63
- Cigarette smoking, lung cancer and, 78-88
- City and County Data Book, 169-70
- Cochran, William G., 3-4
- Coefficients:
 correlation, slope compared to, 101-7
 regression, 76-77
 interpretation of, when variables are re-expressed as logarithms, 108-32
 in multiple regression, 138-39
 scaling of variables and interpretation of, 78
- Coercion by government, 29
- Coleman, James, 153
- Colorado, 23
- Combinations of describing variables, prediction of response variable from, 33-35
- Comparison, controlled, 3-5
Congressional Directory, 93, 169
Congressional District Data Book, 169
 Congressional elections (*see* Elections)
- Congressional hearings, 164-65
Congressional Quarterly, 169
Congressional Record, 165
- Connecticut, traffic accident deaths in, 8, 12, 20, 23
- Conservative party (Great Britain), 95
- Control groups, 3-4
- Controlled comparison, 3-5
- Correlation coefficients, slope compared to, 101-7
- Correlations:
 "nonsense," 88-91
 prediction of, two and three describing variables, 33-35
 spurious, 19, 21, 61-62
- Criminality, prediction of, 36-37
- Crook County (Oregon), 47, 54
- Cross-section analysis in automobile safety inspection study, 7-29
- Cube law, testing with a logit model, 121-32
- Cube-root law in five-variable regression analysis, 156-57
- Dahl, Robert A., 2
- Daniel, Cuthbert, 163 *n*
- Data, research design and collection of, 35
- Deaths (death rate):
 heart disease, 85-88
 lung cancer, smoking and, 78-88
 traffic:
 automobile safety inspections and, 7-29

- per mileage, computation, 16
- population density and, 20-28
- rates, by state, 8-12, 23
- recording of, 14-15
- Definitions, statistical thinking and, 34-35
- Delaware, 23
- Democratic party (U.S.A.):
 - percentage of congressional seats won by (1900-1972), 66-68
 - votes-congressional seats relationship for, 93-95
- Denmark, 82
- Dependent variable, 2
- Describing variables, 2, 18
 - prediction of correlation between two and three, 33-35
- Description, 2
- Design, research:
 - causal analysis and, 3
 - defects in, effects of prior selection, 55-60
 - prediction methods and, 34-35, 55-60
 - regression fallacy in, 56-60
- Deviant cases, 34
- Dichotomous variable (dummy variable), 14, 158-59
- Dodd's cube-root law, 156-57
- Doll, R., 79 *n*
- Dummy variable, 14, 158-59
- Duncan, Otis Dudley, 31

- Economic conditions, midterm congressional elections, presidential popularity and, 139-48
- Educational opportunity, multicollinearity and equality of, 148-55
- Eisenhower, Dwight D., 74
- Elections:
 - forecasting of, 40-55
 - bellwether electoral districts, 46-55
 - interpretation of early returns, 41-55
 - meld projections, 44-46
 - mu curve method, 41-44
 - midterm congressional, presidential popularity, economic conditions and, 139-48
 - 1900-1972, percentage of congressional seats won by Democrats, 66-68
 - 1936, 47
 - 1936-1964, predictive record of bellwether districts, 49-52
 - 1944-1970 congressional, presidential popularity and results of, 73-77
 - 1968, 48-50
 - sources for data on, 169-70
 - swing ratios in congressional, 94-98
 - tabulation of, 93 *n*
- Electoral districts, bellwether, 46-55
- Eliot, T. S., 46
- Errors:
 - random measurement, 57-59
 - standard, of estimate of the slope, 70, 73
- Eskimos, 87
- Estimate of slope (*see* Slope)
- Explanation:
 - causal, 2-5
 - in automobile safety inspections study, 5-29
 - reform and, 2
- Extrapolation, 32-33
 - beyond the data, example, 63-64
 - hidden, 33
- Extreme groups, regression toward the mean by, 56-60

- Fallacy, regression, 56-60
- Farley, James, 47
- Fat calories consumed, heart disease and, 85-88
- Finland, 82
- Fisher, R. A., 148
- Fitted lines:
 - in adjustment method, 25-26
 - measures of quality of, 69-71
 - observed data and, 65-68
 - two-variable linear regression and, 65-73, 78-80
- Florida, 23
- Forecasting, (*see* Predictions)
- Fox, Karl A., 33
- France, 156
- Frequency distributions, logged vs. unlogged, 110-13

- Gallup Poll, 74, 141-43
 Georgia, 23
 Gilbert, John P., 139 *n*
 Governmental bureaucracy, population size and, two variables logged, 119-21
 GPO (U.S. Government Printing Office), 164, 166
 Great Britain, 82-84
 number of radios and mental defectives in, as example of nonsense correlation, 88-91
 relationship between parliamentary seats and votes in, 95, 96
 Grieves, Thomas J., 49
 Guano, 63-64
- Handbook of Political and Social Indicators*, 167
- Hawaii, 8, 23
 Heart disease, mortality from, 85-88
 Hidden extrapolation, 32
 Hiroshima (Japan), 3
 Hirschi, Travis, 36 *n*
 Hobson, Carol J., 153 *n*
 Hodgson, Godfrey, 40 *n*
 Holland, 82
 Holmes, R. A., 60 *n*
 Hudson, Michael C., 167
 Humboldt, Baron, 63
 Hypothesis, causal, 3
- Iceland, 80, 82
 Idaho, 8, 23
 Illinois, 23
 1968 tabulation of votes in, 40
 1970 elections in, 101
 Independent variables, 2
 India, 156
 Indiana, 23
 Inferences, controlled comparison and, 3-4
 Injuries, automobile accident, 5
 Inspections, automobile safety, 5-30
 costs of, 29
 unquantifiable aspects of, 29-30
 Insurance companies, automobile, 60
 Interaction effect, 84
 Intercept:
 least-squares estimate of, 68
 two-variable linear regression and, 80-81
 Interpretation, data collection and, 35
 Inter-University Consortium for Political Research, 47 *n*
 Iowa, 23
 Italy, 86, 156
 Iversen, Gudmund, 91 *n*
- Japan, 86
 gross national product (GNP) of (1961-1970), 126-29
 Johnson, Lyndon B., 75
 Johnston, J., 111 *n*, 153 *n*
- Kansas, 23
 Kelley, Stanley, Jr., 130-32
 Kempthorne, O., 102 *n*
 Kendall, M. G., 88, 90
 Kennedy, John F., 41, 75
 Kentucky, 23
 Key, V. O., 140 *n*
 "Key precincts," 44-45
 Keys, A., 85, 88 *n*
 Kramer, Gerald H., 141 *n*
 Kruskal, Joseph B., 103 *n*, 112 *n*
- Labour party (Great Britain), 95
 Labour party (New Zealand), 95
 Laramie County (Wyoming), 54
 Least squares, principle of, 68
 Least-squares regression, 68-69
 Lecam, Lucien M., 88 *n*
 Legislative seats, relationship between votes and, 91-101
 testing cube law with a logit model, 121-32
 Levin, Harry, 154-55
 Liechtenstein, 156
 Linear regression, two-variable, 65-134
 comparing slope and correlation coefficient, 101-7
 interpretation of regression coefficients when the variables are re-expressed as logarithms, 108-32

- nonsense correlations, 88-91
 President's popularity and results of congressional elections, case study, 73-77
 relationship between seats and votes in two-party systems, case study, 91-101
 smoking and lung cancer, case study, 78-88
 Little, Arthur D., 62 *n*
 Logarithms:
 interpretation of regression coefficients when variables are re-expressed as, 108-32
 review of, 108-13
 Logit model, testing the cube law relating seats and votes with a, 121-32
London Farmers' Magazine: Prospect of the American Guano Company, 64 *n*
 Louisiana, 23
 Lung cancer, smoking and, 78-88
- McPartland, James, 153 *n*
 Maine, 23, 47
 March, James G., 122 *n*
 Markov model, 53
 Maryland, 23
 Massachusetts, 8, 23, 44
 Matching method in causal explanation, 21-22, 24
 Matthew, Douglas, 169
 Maturation, effect on test scores of, 56, 57, 60
 Maugham, Somerset, 54-55
 Mayhew, David R., 99 *n*
 Mean, research design and regression toward the, 56-60
 Meld projections in election forecasting, 44-46
 Michigan, 23, 95, 96
 1970 election in, 100-101
 Minnesota, 23
 Mintz, Morton, 165
 Mirer, Thad W., 130-32
 Mississippi, 23
 Missouri, 23
 Model, probability, for bellwether electoral district, 48, 52-53
 Montana, traffic accident deaths in, 8, 11, 23
 Mood, Alexander M., 153 *n*
 Mortality (*see* Deaths)
 Moshman, Jack, 42 *n*
 Mosteller, Frederick, 17-18, 34-35, 139 *n*, 154 *n*
 Moynihan, Daniel P., 5 *n*, 17-18, 29 *n*, 139 *n*, 154 *n*
 Mu curve, in election forecasting, 41-44
 Mueller, John E., 75 *n*
 Multicollinearity, 148-55
 cures for, 153
 symptoms of, 152
 Multiple regressions, 135-63
 equality of educational opportunity and multicollinearity as case study of, 148-55
 midterm congressional elections as case study of, 139-48
 regression coefficients in, 138-39
 Multivariate analysis, 2, 33
 Myth, faculty for, 54-55
- NACLA Research Methodology Guide*, 165-66
 Nagasaki (Japan), 3
 National Crime Test, 36, 37
 National party (New Zealand), 95
 Natural logarithms, 109
 Nebraska, 23
 Negligent drivers, 62
 Nevada, 8, 20, 23
 New Hampshire, 23
 New Jersey:
 traffic accident deaths in, 5, 8, 20, 23, 24
 votes-congressional seats relationship in, 93, 95, 96
 New Mexico, traffic accident deaths in, 8, 12, 22, 23
 New York, 8, 23, 95, 96
New York Times, The, 17 *n*, 48-49
New York Times Almanac, The, 165, 167
 New Zealand, 95
 Neyman, Jerzy, 88 *n*
 Nixon, Richard Milhous, 75
 Nonsense correlations, 88-91

- North American Congress on Latin America, 165-66
- North Carolina, 23
- North Dakota, 23
- Norway, 82
- Observed relationship, developing explanations for the, 18-29
- Ohio, 23, 101
- Oklahoma, 23
- Oregon, traffic accident deaths in, 22-24
- Page, Bruce, 40 *n*
- Palo Alto County (Iowa), 47, 54
- Parliamentary size:
 five-variable regression analysis of, 156-63
 relationship between population size and, two variables logged, 115-18
- Partial slopes in multiple regression, 138
- Party advantage, 94
- Patterns of association, 3
- Peacock, Dr. E. E., Jr., 4 *n*
- Pennsylvania, 23, 101
- Peru, 63
- Poe, Edgar Allen, 65
- Policy, social, 2
 crude vs. refined measures in study of, 17-18
- Political Handbook and Atlas of the World*, 167
- Polsby, Nelson W., 99 *n*
- Pool, Ithiel de Sola, 31
- Popularity of president:
 economic conditions, midterm congressional elections, 139-48
 results of congressional elections and, 73-77
- Population:
 governmental bureaucracy size and, two variables logged, 119-21
 logged vs. unlogged frequency distribution of, 110-11
 relationship between parliamentary size and, Dodd's cube-root law, 156-57
 relationship between parliamentary size and two variables logged, 115-18
- Population density, death rates from automobile accidents and, 20-28
- Practice, effect on test scores of, 56, 60
- Predictions, 3, 31-64
 in adjustment method, 26-27
 of automobile-accident proneness, 60-62
 case studies in, 35-64
 of correlation between two and three describing variables, 33-35
 of lung-cancer deaths, residual analysis, 81-84
 prior selection as affecting, 55-60
 residuals from, 26-28
- Pregnancy, radiation during, 3, 4
- Presidential popularity:
 economic conditions, midterm congressional elections and, 139-48
 results of congressional elections and, 73-77
- Prior selection, effect of, 55-60
- Probability:
 computation of conditional, Bayes' theorem, 38
 conditional, 37-38
- Probability model for bellwether electoral districts, 48, 52-53
- Projection of election winners, 40-55
 bellwether electoral districts and, 46-55
 meld method of, 44-46
 mu curve method for, 41-44
- Radiation, atomic, 3-4
- Radiological Society of North America, 3-4
- Radios, increase in number of mental defectives and number of (non-sense correlation), 88-91
- Random data, 60
- Random measurement errors added to test scores, 57-59

- Random samples, 6
 Reapportionment, 99
 Redistricting, electoral, 93, 99
 Reform, social, 2, 6
 Regression:
 five-variable, 156-63
 least-squares, 68-69
 multiple, 135-63
 equality of educational opportunity and multicollinearity, case study, 148-55
 midterm congressional elections as case study, 139-48
 regression coefficients, 138-39
 two-variable linear, 65-134
 comparing slope and correlation coefficient, 101-7
 interpretation of regression coefficients when the variables are re-expressed as logarithms, 108-32
 nonsense correlations, 88-91
 President's popularity and results of congressional elections, case study, 73-77
 relationship between seats and votes in two-party system, case study, 91-101
 smoking and lung cancer, case study, 78-88
 Regression coefficients, 76-77
 interpretation of, when variables are re-expressed as logarithms, 108-32
 in multiple regression, 138-39
 scaling of variables and interpretation of, 78
 Regression fallacy, 56-60
 Regression toward the mean, research design and, 56-60
 Replication of bellwether electoral district performance, 48
 Republican National Committee, 168
 Republican Party (U.S.A.), 95
 swing ratios for, 96-98
 Research design:
 causal analysis and, 3
 defects in, effects of prior selection, 55-60
 prediction methods and, 34-35, 55-60
 regression fallacy in, 56-60
 Residual analysis, 81-84
 Residuals, adjustment method and, 26-28
 Residual variation, fitted line and, 69, 76-77
 Response variable, 2, 3
 Rhode Island, 8, 11, 23
 Richards, J. W., 112 *n*
 Roosevelt, Franklin D., 47
 Ross, H. Laurence, 7, 61
 Russett, Bruce, 112 *n*

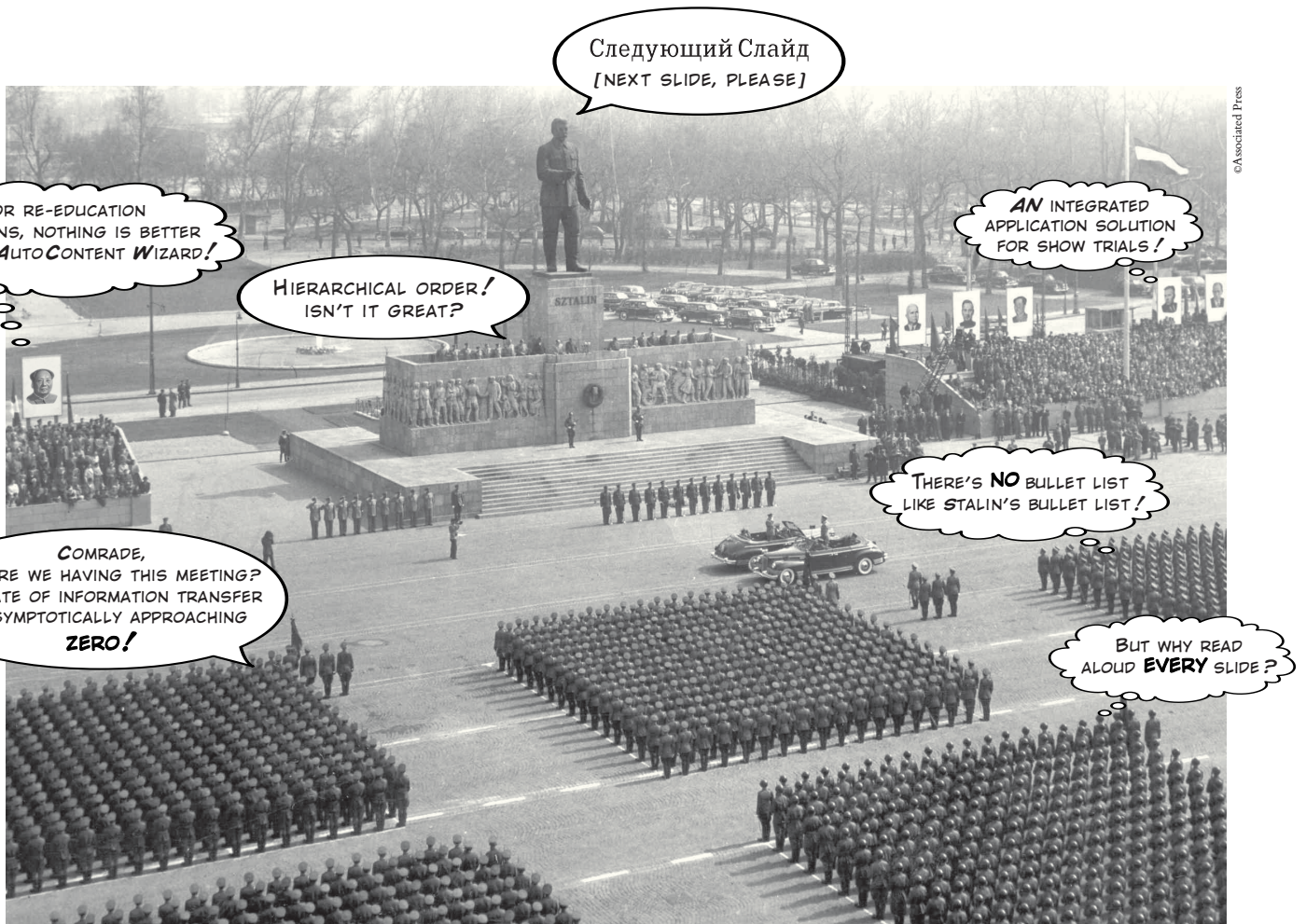
 Safety inspections, automobile, 5-29
 Safety programs, automobile, 7
 Salem (New Jersey), 49
 Samples, random, 6
 San Marino, 156
 Sartwell, P. E., 3, 4 *n*, 126 *n*
 Scaling of variables, interpretation of regression coefficients and, 78
 Scatterplots, two-variable linear regressions and, 132-34
 Schlesinger, Arthur M., 31 *n*
Science (magazine), 166
 Scott, Elizabeth L., 88 *n*
 Seats, relationship between votes and legislative, 91-101
 testing cube law with a logit model, 121-32
 Selection:
 of countries in mortality studies, 85-88
 prior, effects on measuring future performances in tests, 55-60
 self-, 87
 Seltser, R., 3, 4 *n*
 Selvin, Hanan C., 36 *n*
 Skewed variables, logging, 110
 Skolnick, Jerome H., 39 *n*
 Slope:
 correlation coefficient compared to fitted line's, 101-7
 of fitted lines, 66-68, 70
 least-squares estimate of, 68
 partial, in multiple regression, 138
 standard error of estimate of, 70, 73
 statistical significance test for estimate of, 73

- Slope (*cont.*)
 swing ratio estimate and estimate of, 95
- Smoking, lung cancer and, 78-88
- Social policy, 2
 crude vs. refined measures in study of, 17-18
- Social reform, 2, 6
- Somers, Herman, 4 *n*
- Sources of Historical Election Data* (Burnham), 167
- South Carolina, 23
- South Dakota, 23
- Speed, automobile accidents and, 61
- Spurious correlations, 19, 21, 61-62
- Squares, principle of least, 68
- Standard error of the estimate of the slope, 70, 73
- Standardized coefficients, 138
- State almanacs (yearbooks), 167-68
- Statistical Abstract of the United States*, 167
- Statistics:
 thinking vs. formulas in, 34-35
 uses of, 1
- Stigler, George J., 141 *n*
- Stokes, Donald, 91 *n*
- Stone, I. F., 165
- Straight line, equation of a, 66
- Sun, Richard A., 46 *n*
- Survey research, "shotgunning" in, 48
- Sweden, 82, 86, 156
- Swing ratio, 68
 in congressional elections, 94-99
 turnover and, 98-99
- Switzerland, 82
- Taylor, Charles Lewis, 167
- Television networks, 40, 41
- Tennessee, 23
- Test scores, analysis of changes in, 55-60
- Texas, 23
- Thurber, James, 5
- Time-series analysis, 6-7, 153
- Traffic accidents (*see* Accidents, automobile)
- Traffic violations, accident-proneness and, 60-62
- True scores in regression fallacy example, 57-59
- Truman, Harry S., 75
- Tufte, Edward R., 7 *n*, 46 *n*, 65 *n*, 91 *n*, 115 *n*, 121 *n*, 153 *n*
- Tukey, John W., 65 *n*, 101-102
- Turnout at elections, 139-40
- Turnover, swing ratio and, 98-99
- Two-party systems:
 parliamentary size in, 158
 relationship between legislative seats and votes in, 91-101
- Two-variable linear regression (*see* Linear regression, two-variable)
- Ujifusa, Grant, 169
- Unexplained variation, ration of total variation to, 70-72
- United Nations, 167
- United States Government Printing Office (GPO), 164, 166
- Unstandardized coefficients, 138
- Utah, 23
- Variables:
 causal explanations and, 2, 3, 18-20
 clusters of interrelated, 3
 dependent, 2
 describing, 2, 18
 prediction of correlation between two and three, 33-35
 dichotomous (dummy), 14, 158-59
 independent, 2
 response, 2, 3
 scaling of, interpretation of regression coefficients and, 78
 statistical control vs. actual experimental control of, 139
- Variation:
 ratio of explained and unexplained variation to total, 70-72
 residual, 69, 76-77
 sources of, 35
- Vermont, 23, 47
- Viner, Jacob, 147-48
- Virginia, 23
- Volunteers, 6

- Votes, relationship between legislative seats and, 91–101
testing cube law with a logit model, 121–32
- Voting machines, 41, 42
- Wales, 86
- Wallis, W. Allen, 35, 37
- Washington, 23
- Weighted averages in election projections, 45–46
- Weinfield, Frederic, 153 *n*
- West Germany, 156
- West Virginia, 23
- Wilk, M. B., 65 *n*
- “Winsorizing,” 103 *n*
- Winsor, Charles, 108
- Wisconsin, 23
- Wold, Herman, 2
- Wonnacott, Ronald J., 137 *n*
- Wonnacott, Thomas H., 137 *n*
- Wood, Fred, 163 *n*
- World Almanac*, 165, 167
- Wyoming, death rates in automobile accidents in, 8–11, 20, 22–24
- X-rays, 3, 4
- Yerushalmy, J., 85–88
- York, Robert L., 153 *n*
- Yule, G. Udny, 88, 90

Edward R. Tufte

The Cognitive Style of PowerPoint: Pitching Out Corrupts Within



Military parade, Stalin Square, Budapest, April 4, 1956.

Copyright © 2006 by Edward R. Tufte. Second edition. All rights reserved.

Published by Graphics Press LLC P.O. Box 430 Cheshire, Connecticut 06410 www.tufte.com

ISBN 978-1-930824-17-1

The English language . . . becomes ugly and inaccurate because our thoughts are foolish, but the slovenliness of our language makes it easier for us to have foolish thoughts.

George Orwell, "Politics and the English Language"

For a successful technology, reality must take precedence over public relations, for Nature cannot be fooled.

Richard P. Feynman, "*What Do You Care What Other People Think?*"

And not waving but drowning.

Stevie Smith, poem, "Not Waving But Drowning"

Sweet songs never last too long on broken radios.

John Prine, "Sam Stone"

The Cognitive Style of PowerPoint: Pitching Out Corrupts Within

IN corporate and government bureaucracies, the standard method for making a presentation is to talk about a list of points organized onto stylized slides projected up on the wall. For years, before computerized presentations, those giving a talk used transparencies for projected images. Now presenters use a slideware program, Microsoft PowerPoint, which turns out billions and billions of presentation slides each year.

This chapter provides evidence that *compares PowerPoint with alternative methods for presenting information*: 10 case studies, an unbiased collection of 2,000 PP slides, and 32 control samples from non-PP presentations.

The evidence indicates that PowerPoint, compared to other common presentation tools, reduces the analytical quality of serious presentations of evidence. This is especially the case for the PowerPoint ready-made templates, which corrupt statistical reasoning, and often weaken verbal and spatial thinking. What is the problem with PowerPoint? How can we improve our presentations? And what specific sorts of corruptions of evidence and analysis should *consumers* of PowerPoint presentations look out for?

WHEN Louis Gerstner became president of IBM, he encountered a big company caught up in ritualistic slideware-style presentations:

One of the first meetings I asked for was a briefing on the state of the [mainframe computer] business. I remember at least two things about that first meeting with Nick Donofrio, who was then running the System/390 business . . .

At that time, the standard format of any important IBM meeting was a presentation using overhead projectors and graphics that IBMers called “foils” [projected transparencies]. Nick was on his second foil when I stepped to the table and, as politely as I could in front of his team, switched off the projector. After a long moment of awkward silence, I simply said, “Let’s just talk about your business.”

I mention this episode because it had an unintended, but terribly powerful ripple effect. By that afternoon an email about my hitting the Off button on the overhead projector was crisscrossing the world. Talk about consternation! It was as if the President of the United States had banned the use of English at White House meetings.¹

¹ Louis V. Gerstner, Jr., *Who Says Elephants Can’t Dance? Inside IBM’s Historic Turn-around* (2002), 43.

The Cognitive Style of PowerPoint

GERSTNER's blunt action shutting down the projector suggests there are better tools for doing business analysis than reading aloud from bullet lists: "Let's just talk about your business." Indeed, Gerstner later asked IBM executives to write out their business strategies in longhand using the presentation methodology of *sentences*, with subjects and predicates, nouns and verbs, which then combine sequentially to form *paragraphs*, an analytic tool demonstratively better than slideware bullet lists.²

"Let's just talk about your business" indicates a thoughtful exchange of information, a mutual interplay between speaker and audience, rather than a pitch made by a power pointer pointing to bullets. PowerPoint is *presenter-oriented, not content-oriented, not audience-oriented*. PP advertising is not about content quality, but rather presenter therapy: "A cure for the presentation jitters." "Get yourself organized." "Use the AutoContent Wizard to figure out what you want to say."

PowerPoint's convenience for some presenters is costly to the content and the audience. These costs arise from the *cognitive style characteristic of the standard default PP presentation: foreshortening of evidence and thought, low spatial resolution, an intensely hierarchical single-path structure as the model for organizing every type of content, breaking up narratives and data into slides and minimal fragments, rapid temporal sequencing of thin information rather than focused spatial analysis, conspicuous chartjunk and PP Phluff, branding of slides with logotypes, a preoccupation with format not content, incompetent designs for data graphics and tables, and a smirky commercialism that turns information into a sales pitch and presenters into marketeers*. This cognitive style harms the quality of thought for the producers and the consumers of presentations.

PowerPoint comes with a big attitude. Other than video games, not many computer programs have attitudes. Effective tools such as web browsers, Word, Excel, Photoshop, and Illustrator are not accompanied by distinctive cognitive styles that reduce the intellectual level of the content passing through the program.

Nonetheless, PowerPoint may benefit the bottom 10% of all presenters. PP forces them to have *points*, some points, any points. Slideware perhaps helps inept speakers get their act together, outline talks, retrieve visual materials, present slides. Furthermore, PP probably doesn't cause much damage to really first-rate presenters, say the top 10%, who have strong content, self-awareness, and their own analytical style that avoids or neutralizes the PP style. This leaves 80%, workaday presenters, for whom the PP cognitive style causes trouble.

In practice, PP slides are very low resolution compared to paper, most computer screens, and the immense visual capacities of the human eye-brain system. With little information per slide, many many slides are needed. Audiences endure a relentless sequentiality, one damn slide after

² Gordon Shaw, Robert Brown, Philip Bromiley, "Strategic Stories: How 3M is Rewriting Business Planning," *Harvard Business Review*, 76 (May-June, 1998), 42-44.



another. Information stacked in time makes it difficult to understand context and evaluate relationships. Visual reasoning usually works more effectively when the relevant evidence is shown *adjacent in space* within our eyespan. This is especially the case for statistical data, where the fundamental analytical task is to make comparisons.

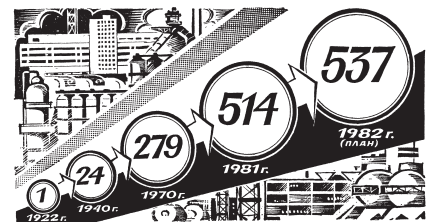
The statistical graphics produced by PowerPoint are astonishingly thin, nearly content-free. In 28 books on PP templates, the 217 model statistical graphics depict an average of 12 numbers each (as do the PP data-table templates). Compared to the worldwide publications shown here, the PP statistical graphics are the thinnest of all, except for those in *Pravda* in 1982, back when that newspaper operated as the major propaganda instrument of the Soviet communist party and a totalitarian government.³ Doing a bit better than *Pravda* is not good enough:

MEDIAN NUMBER OF ENTRIES IN DATA MATRICES FOR STATISTICAL GRAPHICS IN VARIOUS PUBLICATIONS, 2003

<i>Science</i>	> 1,000
<i>Nature</i>	> 700
<i>New York Times</i>	120
<i>Wall Street Journal</i>	112
<i>Frankfurter Allgemeine Zeitung</i>	98
<i>New England Journal of Medicine</i>	53
<i>Asahi</i>	40
<i>Financial Times</i>	40
<i>The Economist</i>	32
<i>Le Monde</i>	28
28 books on PowerPoint presentations (1997-2003)	12
<i>Pravda</i> (1982)	5



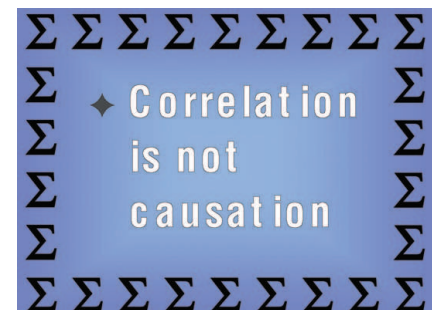
³ In this table, the medians are based on at least 20 statistical graphics and at least one full issue of each publication. These publications, except for scientific journals, tend to use the same graph designs issue after issue; thus replications of several of the counts were within 10% of the original result. Data for other publications (*Pravda*, for example) are reported in Edward R. Tufte, *The Visual Display of Quantitative Information* (1983, 2001), 167.



Pravda, May 24, 1982.

These PP graph templates are particularly unfortunate for students, since for all too many their *first* experience in presenting statistical evidence is via PP designs, which create the impression that data graphics are for propaganda and advertisements and not for reasoning about information.

And, in presenting *words*, impoverished space encourages imprecise statements, slogans, abrupt and thinly-argued claims. For example, this slide from a statistics course shows a seriously incomplete cliché. In fact, probably the *shortest true statement* that can be made about causality and correlation is “*Empirically observed covariation is a necessary but not sufficient condition for causality.*” Or perhaps “*Correlation is not causation but it sure is a hint.*” Many true statements are too long to fit on a PP slide, but this does not mean we should abbreviate the truth to make the words fit. It means we should find a better tool to make presentations.



Sequentiality of the Slide Format

WITH information quickly appearing and disappearing, the slide transition is an event that attracts attention to the presentation's compositional methods. Slides serve up small chunks of promptly vanishing information in a restless one-way sequence. It is not a contemplative analytical method; it is like television, or a movie with over-frequent random jump cuts. Sometimes quick chunks of thin data may be useful (flash-card memorizing), other times not (comparisons, links, explanations). *But formats, sequencing, and cognitive approach should be decided by the character of the content and what is to be explained, not by the limitations of the presentation technology.* The talk that accompanies PP slides may overcome the noise and clutter that results from slideville's arbitrary partitioning of data, but why disrupt the signal in the first place? And why should we need to recover from a technology that is supposed to help our presentations?

Obnoxious transitions and partitions occur not only slide-by-slide but also line-by-line, as in the dreaded slow reveal (at right). Beginning with a title slide, the presenter unveils and reads aloud the single line on the slide, then reveals the next line, reads that aloud, on and on, as the stupefied audience impatiently awaits the end of the talk.

It is helpful to provide audience members with at least one mode of information that allows *them* to control the order and pace of learning—unlike slides and unlike talk. Paper handouts for talks will help provide a permanent record for review—again unlike projected images and talk. Another way to break free of low-resolution temporal comparisons is to show multiple slides, several images at once within the common view. Spatial parallelism takes advantage of our notable capacity to reason about multiple images that appear simultaneously within our eyespan. We are able to select, sort, edit, reconnoiter, review—ways of seeing quickened and sharpened by direct spatial adjacency of evidence.

Now and then the narrow bandwidth and relentless sequencing of PP slides are said to be virtues, a claim justified by loose reference to George Miller's classic 1956 paper "The Magical Number Seven, Plus or Minus Two." That essay reviews psychological experiments that discovered people had a hard time remembering more than about 7 unrelated pieces of really dull data all at once. These studies on memorizing nonsense then led some interface designers, as well as PP guideline writers seeking to make a virtue of a necessity, to conclude that only 7 items belong on a list or a slide, a conclusion that can only be reached by not reading Miller's paper. In fact the paper neither states nor implies rules for the amount of information shown on a slide (except for those presentations consisting of nonsense syllables that the audience must memorize and repeat back to a psychologist). On the contrary, the deep point of Miller's work is to suggest strategies, such as placing evidence within a context, that extend the reach of memory beyond tiny clumps of data.⁴

The Dreaded Build Sequence

The Dreaded Build Sequence

THE FIRST LINE IS REVEALED

The Dreaded Build Sequence

THE FIRST LINE IS REVEALED

THE SECOND LINE IS
REVEALED!

The Dreaded Build Sequence

THE FIRST LINE IS REVEALED

THE SECOND LINE IS
REVEALED!

THE THIRD LINE IS REVEALED

[THE AUDIENCE FLEES]

⁴ George A. Miller, "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information," *Psychological Review*, 63 (1956), 81-97 (and widely posted on the internet). At Williams College in September 2000, I saw George Miller give a presentation that used the optimal number of bullet points on the optimal number of slides—zero in both cases. Just a straightforward talk with a long narrative structure.

Metaphors for Presentations and Conway's Law

THE metaphor of PowerPoint is *the software corporation itself*. To describe a software house is to describe the PP cognitive style: a big bureaucracy engaged in *computer programming* (deep hierarchical structures, relentlessly sequential, nested, one-short-line-at-a-time) and in *marketing* (advocacy not analysis, more style than substance, misdirection, slogan thinking, fast pace, branding, exaggerated claims, marketplace ethics). That the PP cognitive style mimics a software house exemplifies *Conway's Law*:

Any organization which designs a system . . . will inevitably produce a design whose structure is a copy of the organization's communication structure.⁵

Why should the structure, activities, and values of a large commercial bureaucracy be a useful metaphor for our presentations? Are there worse metaphors? Voice-mail menu systems? Billboards? Television? Stalin?

The pushy PP style tends to set up a dominance relationship between speaker and audience, as the speaker makes power points with hierarchical bullets to passive followers. Such aggressive, stereotyped, over-managed presentations—the Great Leader up on the pedestal—are characteristic of hegemonic systems *and of Conway's Law again in operation*:

The Roman state bolstered its authority and legitimacy with the trappings of ceremony. . . . Power is a far more complex and mysterious quality than any apparently simple manifestation of it would appear. It is as much a matter of impression, of theatre, of persuading those over whom authority is wielded to collude in their subjugation. Insofar as power is a matter of presentation, its cultural currency in antiquity (and still today) was the creation, manipulation, and display of images. In the propagation of the imperial office, at any rate, art was power.⁶

A BETTER metaphor for presentations is *good teaching*. Practical teaching techniques are very helpful for presentations in general. Teachers seek to explain something with credibility, which is what many presentations are trying to do. The core ideas of teaching—*explanation, reasoning, finding things out, questioning, content, evidence, credible authority not patronizing authoritarianism*—are contrary to the cognitive style of PowerPoint. And the ethical values of teachers differ from those engaged in marketing.⁷

Especially disturbing is the introduction of PowerPoint into schools. Instead of writing a report using sentences, children learn how to decorate client pitches and infomercials, which is better than encouraging children to smoke. Student PP exercises (as seen in teachers' guides, and in student work posted on the internet) typically show 5 to 20 words and a piece of clip art on each slide in a presentation consisting of 3 to 6 slides—a total of perhaps 80 words (20 seconds of silent reading) for a week of work. Rather than being trained as mini-bureaucrats in the pitch culture, students would be better off if schools closed down on PP days and everyone went to The Exploratorium. Or wrote an illustrated essay explaining something.

⁵ Melvin E. Conway, "How Do Committees Invent?," *Datamation*, April 1968, 28–31. The law's "inevitably" overreaches. Frederick P. Brooks, Jr., in *The Mythical Man-Month: Essays on Software Engineering* (1975), famously describes the interplay between system design and bureaucracy.

⁶ Jás Elsner, *Imperial Rome and Christian Triumph: The Art of the Roman Empire AD 100–450* (Oxford, 1998), 53.

⁷ On teaching, see Joseph Lowman, *Mastering the Techniques of Teaching* (San Francisco, 1995); Wilbert McKeachie and Barbara K. Hofer, *McKeachie's Teaching Tips* (New York, 2001); Frederick Mosteller, "Classroom and Platform Performance," *The American Statistician*, 34 (1980), 11–17 (posted at www.edwardtufte.com).

PowerPoint Does Rocket Science: Assessing the Quality and Credibility of Technical Reports

NEARLY all engineering presentations at NASA are made in PowerPoint. Is this a product endorsement or a big mistake? Does PP's cognitive style affect the quality of engineering analysis? How does PP compare with alternative methods of technical presentation? Some answers come from the evidence of NASA PowerPoint in action: (1) hundreds of PP technical presentations experienced in 2003 by the Columbia Accident Investigation Board and in 2005 by the Return to Flight Task Group, (2) a case study of the PP presentations for NASA officials making life-and-death decisions during the final flight of Columbia, (3) observations by Richard Feynman who saw a lot of slideware-style presentations in his NASA work on the 1986 Challenger accident, (4) my observations as a NASA consultant on technical presentations for shuttle risk assessments, shuttle engineering, and deep spaceflight trajectories.

DURING the January 2003 spaceflight of shuttle Columbia, 82 seconds after liftoff, a 1.67 pound (760 grams) piece of foam insulation broke off from the liquid fuel tank, hit the left wing, and broke through the wing's thermal protection. After orbiting the Earth for 2 weeks with an undetected hole in its wing, Columbia burned up during re-entry because the compromised thermal protection was unable to withstand the intense temperatures that occur upon atmosphere re-entry. The 7 astronauts on board died. The only evidence of a possible problem was a brief video sequence showing that something hit the wing somewhere. Here are 2 video frame-captures at 82 seconds after Columbia's launch:



The rapidly accelerating Columbia in effect ran into the foam debris. Post-accident frame-by-frame analysis yields the impact velocity of the foam, 600 miles or 970 km per hour, the speed of sound. Since kinetic energy = $\frac{1}{2}mv^2$, the velocity-squared contribution is substantial.

In the video, 2 relevant variables are indeterminate: impact *angle of incidence* and impact *location*. Did the debris hit the insulation tiles on the left wing, or the reinforced carbon-carbon (RCC) on the leading edge of the wing? Post-accident investigation established that the foam hit the especially vulnerable RCC.

What to make of this video? How serious is the threat? What actions should be taken in response? A quick, smart analysis is needed, since Columbia will re-enter the atmosphere in about 12 days. Although the evidence is uncertain and thin, for only a single camera showed debris impact, the logical structure of the engineering analysis is straightforward:


debris *kinetic energy* (function of mass, velocity, and angle of incidence) + debris hits locations of *varying vulnerability* on left wing → *level of threat* to the Columbia during re-entry heating of wing

Angle of incidence is uncertain; *location of impact* is uncertain (wing tiles? leading edge of the wing?); *mass* and *velocity* of the foam debris can be calculated. Profoundly relevant is the *difference in velocity* between the shuttle and the piece of free-floating foam, since the kinetic energy of the foam impact is proportional to that *velocity squared*. Even though the errant foam was lightweight (1.67 lb), it was moving fast (600 mph) relative to the shuttle. Velocity squared is like shipping and handling: it will get you every time.

To help NASA officials assess the threat, Boeing Corporation engineers quickly prepared 3 reports, a total of 28 PowerPoint slides, dealing with the debris impact.⁸ These reports provided mixed readings of the threat to the spacecraft; the lower-level bullets often mentioned doubts and uncertainties, but the highlighted executive summaries and big-bullet conclusions were quite optimistic. Convinced that the reports indicated no problem rather than uncertain knowledge, high-level NASA officials decided that the Columbia was safe and, furthermore, that no additional investigations were necessary. Several NASA engineers had hoped that the military would photograph the shuttle in orbit with high-resolution spy cameras, which would have easily detected the damage, but even that checkup was thought unnecessary given the optimism of the 3 Boeing reports. And so the Columbia orbited for 16 days with a big undetected hole in its wing.

ON the next page, I examine a key slide in the PP reports made while Columbia was damaged but still flying. The analysis suggests methods for how not to get fooled while consuming a presentation. Imagine that you are a high-level NASA decision-maker receiving a pitch about threats to the spacecraft. You must learn 2 things: Exactly what is the presenter's story? And, can you *believe* the presenter's story? A close reading of a presentation will help gauge the quality of intellect, the knowledge, and the credibility of presenters. To be effective, close readings must be based on *universal* standards of evidence quality, which are not necessarily those standards that operate locally.

⁸ C. Ortiz, A. Green, J. McClymonds, J. Stone, A. Khodadoust, "Preliminary Debris Transport Assessment of Debris Impacting Orbiter Lower Surface in STS-107 Mission," January 21, 2003; P. Parker, D. Chao, I. Norman, M. Dunham, "Orbiter Assessment of STS-107 ET Bipod Insulation Ramp Impact," January 23, 2003; C. Ortiz, "Debris Transport Assessment of Debris Impacting Orbiter Lower Surface in STS-107 Mission," January 24, 2003. These reports were published in records of the CAIB and at NASA websites.

Summary and Conclusion
<ul style="list-style-type: none"> ● Impact analysis ("Crater") indicates potential for large TPS damage <ul style="list-style-type: none"> – Review of test data shows wide variation in impact response – RCC damage limited to coating based on soft SOFI ● Thermal analysis of wing with missing tile is in work <ul style="list-style-type: none"> – Single tile missing shows local structural damage is possible, but no burn through – Multiple tile missing analysis is on-going ● M/OD criteria used to assess structural impacts of tile loss <ul style="list-style-type: none"> – Allows significant temperature exceedance, even some burn through <ul style="list-style-type: none"> • Impact to vehicle turnaround possible, but maintains safe return capability
<p>Conclusion</p> <ul style="list-style-type: none"> ● Contingent on multiple tile loss thermal analysis showing no violation of M/OD criteria, safe return indicated even with significant tile damage

13


The Very Big Bullet phrase fragment does not seem to make sense. No other VBBs appear in the rest of the slide, so this VBB is not necessary.

Spray On Foam Insulation, a fragment of which caused the hole in the wing

A model to estimate damage to the tiles protecting flat surfaces of the wing

Review of Test Data Indicates Conservatism for Tile Penetration

- **The existing SOFI on tile test data used to create Crater was reviewed along with STS-87 Southwest Research data**
 - **Crater overpredicted penetration of tile coating significantly**
 - ◆ **Initial penetration to described by normal velocity**
 - Varies with volume/mass of projectile (e.g., 200ft/sec for 3cu. In)
 - ◆ **Significant energy is required for the softer SOFI particle to penetrate the relatively hard tile coating**
 - Test results do show that it is possible at sufficient mass and velocity
 - ◆ **Conversely, once tile is penetrated SOFI can cause significant damage**
 - Minor variations in total energy (above penetration level) can cause significant tile damage
 - **Flight condition is significantly outside of test database**
 - ◆ **Volume of ramp is 1920cu in vs 3 cu in for test**



On this one Columbia slide, a PowerPoint festival of bureaucratic hyper-rationalism, 6 different levels of hierarchy are used to display, classify, and arrange 11 phrases:

- Level 1 Title of Slide
- Level 2 ● Very Big Bullet
- Level 3 – big dash
- Level 4 ◆ medium-small diamond
- Level 5 • tiny bullet
- Level 6 () parentheses ending level 5


This slide begins with the dreaded Executive Summary, a conclusion presented as a headline: “Test Data Indicates Conservatism for Tile Penetration.” This turns out to be unmerited reassurance. Executives, at least those who don’t want to get fooled, had better read far beyond the title.

The “conservatism” concerns the *choice of models* used to predict damage. But why, after 112 flights, are foam-debris models being calibrated during a crisis? How can “conservatism” be inferred from a loose comparison of a spreadsheet model and some thin data? Divergent evidence means divergent evidence, not inferential security. Claims of analytic “conservatism” should be viewed with skepticism by presentation consumers. Such claims are often a rhetorical tactic that substitutes verbal fudge factors for quantitative assessments.

Here “ramp” refers to foam debris (from the bipod ramp) that hit Columbia. Instead of the cryptic “Volume of ramp,” say “estimated volume of foam debris that hit the wing.” Such clarifying phrases, which may help upper level executives understand what is going on, are too long to fit on low-resolution bullet outline formats. PP demands a shorthand of acronyms, phrase fragments, clipped jargon, and vague pronoun references in order to get at least some information into the tight format.

Review of Test Data Indicates Conservatism for Tile Penetration

- The existing SOFI on tile test data used to create Crater was reviewed along with STS-87 Southwest Research data
 - Crater overpredicted penetration of tile coating **significantly**
 - ◆ Initial penetration to described by normal velocity
 - Varies with volume/mass of projectile (e.g., 200ft/sec for 3cu. In)
 - ◆ **Significant** energy is required for the softer SOFI particle to penetrate the relatively hard tile coating
 - Test results do show that it is possible at sufficient mass and velocity
 - ◆ Conversely, once tile is penetrated SOFI can cause **significant** damage
 - Minor variations in total energy (above penetration level) can cause **significant** tile damage
 - Flight condition is **significantly** outside of test database
 - ◆ Volume of ramp is 1920cu in vs 3 cu in for test



What does this mean?

As the bullet points march on, the seemingly reassuring headline fades away. Lower-level bullets at the end of the slide undermine the executive summary. This third-level point notes that “Flight condition [that is, the debris hit on the Columbia] is significantly outside of test database.” How far outside? The final bullet will tell us.

This fourth-level bullet concluding the slide reports that the debris hitting the Columbia is estimated to be $1920/3 = 640$ times larger than data used in the tests of the model! The correct headline should be “Review of Test Data Indicates Irrelevance of Two Models.” This is a powerful conclusion, indicating that pre-launch safety standards no longer hold. The original optimistic headline has been eviscerated by the lower-level bullets. Note how close attentive readings can help consumers of presentations evaluate the presenter’s reasoning and credibility.

The vigorous but vaguely quantitative words “**significant**” and “**significantly**” are used five times on this slide, with meanings ranging from “detectable in a perhaps irrelevant calibration case study” to “an amount of damage so that everyone dies” to “a difference of 640-fold.” The five “significants” cannot refer to statistical significance, for no formal statistical analysis has been done.

Note the analysis is about *tile* penetration. But what about RCC penetration? As investigators later demonstrated, the foam did not hit the tiles on the wing surface, but instead the delicate reinforced-carbon-carbon (RCC) protecting the wing leading edge. Alert consumers should carefully watch how presenters delineate *the scope of their analysis*, a profound and sometimes decisive matter.

Review of Test Data Indicates Conservatism for Tile Penetration

- The existing SOFI on tile test data used to create Crater was reviewed along with STS-87 Southwest Research data
 - Crater overpredicted penetration of tile coating **significantly**
 - ◆ Initial penetration to described by normal velocity
 - Varies with volume/mass of projectile (e.g., 200ft/sec for 3cu. In)
 - ◆ **Significant** energy is required for the softer SOFI particle to penetrate the relatively hard tile coating
 - Test results do show that it is possible at sufficient mass and velocity
 - ◆ Conversely, once tile is penetrated SOFI can cause **significant** damage
 - Minor variations in total energy (above penetration level) can cause **significant** tile damage
 - Flight condition is **significantly** outside of test database
 - ◆ Volume of ramp is 1920cu in vs 3 cu in for test



Slideville's low resolution and large type generate space-wasting typographic orphans, lonely words dangling on 4 separate lines:

Penetration **significantly** 3cu. In and velocity

The really vague pronoun reference "it" refers to *damage to the left wing*, which ultimately destroyed Columbia (although the slide here deals with tile, not RCC damage). Low-resolution presentation formats encourage vague references because there isn't enough space for specific and precise phrases.

The same unit of measurement for volume (cubic inches) is shown in a different way every time

3cu. In **1920cu in** **3 cu in**

rather than in clear and tidy exponential form 1920 in^3 . Shakiness in conventions for units of measurement should always provoke concern, just as it does in grading the problem sets of sophomore engineering students.* PowerPoint is not good at math and science; here at NASA, engineers are using a presentation tool that makes it difficult to write scientific notation. The pitch-style typography of PP is hopeless for science and engineering, yet this important analysis relied on PP. Technical reports in real science and engineering are not published in PP; how then can PP be used for any serious technical analysis, such as diagnosing the threat to Columbia?

*The Columbia Accident Investigation Board (final report, p. 191) referred to this point about units of measurement: "While such inconsistencies might seem minor, in highly technical fields like aerospace engineering a misplaced decimal point or mistaken unit of measurement can easily engender inconsistencies and inaccuracies." The phrase "mistaken unit of measurement" is an unkind veiled reference to a government agency that had crashed \$250 million of spacecraft into Mars because of a mix-up between metric and non-metric units of measurement.

In the reports, *every single text-slide* uses bullet-outlines with 4 to 6 levels of hierarchy. Then another multi-level list, another bureaucracy of bullets, *starts afresh* for a new slide. How is it that each elaborate architecture of thought always fits *exactly* on one slide? The rigid slide-by-slide hierarchies, indifferent to content, slice and dice the evidence into arbitrary compartments, producing an anti-narrative with choppy continuity. Medieval in its preoccupation with hierarchical distinctions, the PowerPoint format signals every bullet's status in 4 or 5 different simultaneous ways: by the order in sequence, extent of indent, size of bullet, style of bullet, and size of type associated with various bullets. This is a lot of insecure format for a simple engineering problem. The format reflects a common conceptual error in analytic design: information architectures mimic the hierarchical structure of large bureaucracies pitching the information. Conway's Law again. In their report, the Columbia Accident Investigation Board (CAIB) found that the distinctive cognitive style of PowerPoint interacted with the biases and hierarchical filtering of the bureaucracy during the crucial period when the spacecraft was damaged but still functioning:

The Mission Management Team Chair's position in the hierarchy governed what information she would or would not receive. Information was lost as it traveled up the hierarchy. A demoralized Debris Assessment Team did not include a slide about the need for better imagery in their presentation to the Mission Evaluation Room. Their presentation included the Crater analysis, which they reported as incomplete and uncertain. However, the Mission Evaluation Room manager perceived the Boeing analysis as rigorous and quantitative. The choice of headings, arrangement of information, and size of bullets on the key chart served to highlight what management already believed. The uncertainties and assumptions that signaled danger dropped out of the information chain when the Mission Evaluation Room manager condensed the Debris Assessment Team's formal presentation to an informal verbal brief at the Mission Management Team meeting.⁹

⁹ Columbia Accident Investigation Board, *Report*, volume 1 (August 2003), 201.

At about the same time, lower-level NASA engineers were writing about possible dangers to Columbia in several hundred emails, with the Boeing reports in PP format sometimes attached. The text of about 90% of these emails simply used *sentences* sequentially ordered into *paragraphs*; 10% used bullet lists with 2 or 3 levels. These engineers were able to reason about the issues without employing the endless hierarchical outlines of the original PP pitches. Good for them.

Several of these emails referred to the 3 PP reports as the "Boeing PowerPoint Pitch." This is astonishing language. The WhatPoint Pitch? The PowerWhat Pitch? The PowerPoint What? *The language, attitude, and presentation tool of the pitch culture had penetrated throughout the NASA organization, even into the most serious technical work, a real-time engineering analysis of threats to the survival of the shuttle.*

The analysis of the key Columbia slide on the preceding pages was posted at my website.¹⁰ Much of this material was then later included in the final report of Columbia Accident Investigation Board. In their discussion of “Engineering by Viewgraphs,” the Board went far beyond my case study of the Columbia slide in these extraordinary remarks about PowerPoint:

As information gets passed up an organization hierarchy, from people who do analysis to mid-level managers to high-level leadership, key explanations and supporting information are filtered out. In this context, it is easy to understand how a senior manager might read this PowerPoint slide and not realize that it addresses a life-threatening situation.

At many points during its investigation, the Board was surprised to receive similar presentation slides from NASA officials in place of technical reports. The Board views the endemic use of PowerPoint briefing slides instead of technical papers as an illustration of the problematic methods of technical communication at NASA.¹¹

The Board makes an explicit comparison: some tools are better than others for engineering, and technical reports are better than PowerPoint.

THEN, 2 years later, 7 members of the Return to Flight Task Group, a powerful external review group created by NASA to monitor the post-Columbia repairs of the shuttle, had something to say about engineering by PowerPoint. After seeing hundreds of PP decks from NASA and its contractors, the Task Group made direct comparisons of alternative presentation tools for engineering analysis and documentation:

We also observed that instead of concise engineering reports, decisions and their associated rationale are often contained solely within Microsoft PowerPoint charts or emails. The CAIB report (vol. 1, pp. 182 and 191) criticized the use of PowerPoint as an engineering tool, and other professional organizations have also noted the increased use of this presentation software as a substitute for technical reports and other meaningful documentation. PowerPoint (and similar products by other vendors), as a method to provide talking points and present limited data to assembled groups, has its place in the engineering community; however, these presentations should never be allowed to replace, or even supplement, formal documentation.

Several members of the Task Group noted, as had CAIB before them, that many of the engineering packages brought before formal control boards were documented *only* in PowerPoint presentations. In some instances, requirements are defined in presentations, approved with a cover letter, and never transferred to formal documentation. Similarly, in many instances when data was requested by the Task Group, a PowerPoint presentation would be delivered without supporting engineering documentation. It appears that many young engineers do not understand the need for, or know how to prepare, formal engineering documents such as reports, white papers, or analyses.¹²

¹⁰ “Columbia Evidence—Analysis of Key Slide,” March 18, 2003, Ask E.T. forum, www.edwardtufte.com

¹¹ Columbia Accident Investigation Board, *Report*, vol. 1 (August 2003), 191.

¹² Dan L. Crippen, Charles C. Daniel, Amy K. Donahue, Susan J. Helms, Susan Morrissey Livingstone, Rosemary O’Leary, William Wegner, “A.2, Observations,” in *Final Report of the Return to Flight Task Group* (July 2005), 190.

The Return to Flight Task Group made their evaluations and decisions based on closure packages that described the post-Columbia shuttle repairs. In the final report, 7 Task Group members reported that these “inadequate and disorganized” packages, often huge decks of PP slides, provoked “our frustration.”¹³

Closure packages, which should have represented the auditable, documented status of the NASA implementation of the CAIB recommendations, tended to rely on mass, rather than accuracy, as proof of closure. The closure packages showed an organization that apparently still believes PowerPoint presentations adequately explain work and document accomplishments.¹⁴

In an example of the pitch culture in action, some closure packages were provided prematurely to the Return to Flight Task Group in apparent behind-the-scenes maneuvers to discover just what it might take to get approval for the post-accident shuttle repairs. The idea might have been that if it is too late to change the engineering, then change the pitch about the engineering. The Task Group thus found it necessary to repeat Richard Feynman’s famous conclusion to his report on the first shuttle accident, the 1986 loss of the Challenger: “For a successful technology, reality must take precedence over public relations, for Nature cannot be fooled.”¹⁵

By using PP to report technical work, presenters quickly damage their credibility—as was the case for NASA administrators and engineers pitching their usual PP decks to these 2 very serious review boards.

Both the Columbia Accident Investigation Board and the Return to Flight Task Group were filled with smart experienced people with spectacular credentials. These review boards examined what is probably the best evidence available on PP for technical work: hundreds of PP decks from a high-IQ government agency thoroughly practiced in PP. Both review boards concluded that (1) PowerPoint is an inappropriate tool for engineering reports, presentations, documentation and (2) the technical report is superior to PP. Matched up against alternative tools, PowerPoint lost.

Serious problems require a serious tool: written reports. For nearly all engineering and scientific communication, instead of PowerPoint, *the presentation and reporting software should be a word-processing program* capable of capturing, editing, and publishing text, tables, data graphics, images, and scientific notation. Replacing PowerPoint with Microsoft Word (or, better, a tool with non-proprietary universal formats) will make presentations and their audiences smarter. Of course full-screen projected images and videos are necessary; that is the one harmless use of PP. Meetings should center on concisely written reports on paper, not fragmented bulleted talking points projected up on the wall. A good model for the technical report is a scientific paper or commentary on a paper published in substantial scientific journals such as *Nature* or *Science*.

¹³ *Final Report of the Return to Flight Task Group* (July 2005), 195.

¹⁴ *Final Report of the Return to Flight Task Group* (July 2005), 195.

¹⁵ Richard P. Feynman, *What Do You Care What Other People Think? Further Adventures of a Curious Character* (New York, 1988), 237; and quoted by the *Final Report of the Return to Flight Task Group* (July 2005), 194.

High-Resolution Visual Channels Are Compromised by PowerPoint

A TALK, which proceeds at a pace of 100 to 160 spoken words per minute, is not an especially high-resolution method of data transmission. Rates of transmitting *visual* evidence can be far higher. The artist Ad Reinhardt said, “As for a picture, if it isn’t worth a thousand words, the hell with it.” People can quickly look over tables with hundreds of numbers in the financial or sports pages in newspapers. People read 300 to 1,000 printed words a minute, and find their way around a printed map or a 35 mm slide displaying 5 to 40 MB in the visual field. Often the visual channel is an intensely high-resolution channel.

Yet, in a strange reversal, nearly all PowerPoint slides that accompany talks have much *lower* rates of information transmission than the talk itself. Too often the images are content-free clip art, the statistical graphics don’t show data, and the text is grossly impoverished. As shown in this table, *the PowerPoint slide typically shows 40 words, which is about 8 seconds of silent reading material*. The example slides in PP textbooks are particularly disturbing: in 28 books, which should use first-rate examples, the median number of words per slide is 15, worthy of billboards, about 3 or 4 seconds of silent reading material.

This poverty of content has several sources. *The PP design style*, which uses about 40% to 60% of the space available on a slide to show unique content, with remaining space devoted to Phluff, bullets, frames, and branding. *The slide projection of text*, which requires very large type so the audience can see the words. Most importantly, *presenters who don’t have all that much to say* (for example, among the 2,140 slides reported in this table, the really lightweight slides are found in the presentations made by educational administrators and their PR staff).

A vicious circle results. Thin content leads to boring presentations. To make them unboring, PP Phluff is added, damaging the content, making the presentation even more boring, requiring more Phluff . . .

What to do? For serious presentations, it will be useful to replace PowerPoint slides with paper handouts showing words, numbers, data graphics, images together. High-resolution handouts allow viewers to contextualize, compare, narrate, and recast evidence. In contrast, data-thin, forgetful displays tend to make audiences ignorant and passive, and also to diminish the credibility of the presenter. Thin visual content prompts suspicions: “What are they leaving out? Is that all they know? Does the speaker think we’re stupid?” “What are they hiding?” Sometimes PowerPoint’s low resolution is said to promote a clarity of reading and thinking. Yet in visual reasoning, art, typography, cartography, even sculpture, *the quantity of detail is an issue completely separate from the difficulty of reading*.¹⁶ Indeed, quite often, the more intense the detail, the *greater* the clarity and understanding—because meaning and reasoning are relentlessly *contextual*. Less is a bore.

WORDS ON TEXT-ONLY POWERPOINT SLIDES

26 slides in the 3 Columbia reports by Boeing, median number of words per slide	97
1,460 text-only slides in 189 PP reports posted on the internet and top-ranked by Google, March 2003, median number of words per slide	40
654 slides in 28 PowerPoint textbooks, published 1997–2003, median number of words per slide	15

¹⁶ Edward Tufte, *Envisioning Information* (Cheshire, Connecticut, 1990), 36–51.

Sentences Are Smarter Than The Grunts of Bullet Lists

LISTS often serve well for prompts, reminders, outlines, filing, and possibly for quick no-fooling-around messages. Lists have diverse architectures: elaborately ordered to disordered, linearly sequential to drifting in 2-space, and highly calibrated hierarchies of typographic dingbats to free-wheeling dingbat dingbats. In the construction of lists, a certain convenience derives from their lack of syntactic and intellectual discipline, as each element simply consists of scattered words in fragmented pre-sentence grunts.

PowerPoint promotes the hierarchical bullet list, as exemplified in the Columbia slides. The hierarchical bullet list is surely the most widely used format in corporate and government presentations. Slides are filled with over-twiddly structures with some space left over for content. Sometimes the hierarchies are so complex and intensely nested that they resemble computer code, a lousy metaphor for presentations. These formats usually require deeply indented lines for elements consisting of a few words, the power points. The more elaborate the hierarchy, the greater the loss of explanatory resolution, as the container dominates the thing contained.

It is thoughtless and arrogant to replace the sentence as the basic unit for explaining something. Especially as the byproduct of some marketing presentation software.

For the naive, bullet lists may create the appearance of hard-headed organized thought. But in the reality of day-to-day practice, the PP cognitive style is faux-analytical, with a bias towards promoting effects without causes. A study in the *Harvard Business Review* found generic, superficial, simplistic thinking in bullet lists widely used in business planning and corporate strategy:

In every company we know, planning follows the standard format of the bullet outline. . . [But] bullet lists encourage us to be lazy . . .

Bullet lists are typically too generic. They offer a series of things to do that could apply to any business. . . .

Bullets leave critical relationships unspecified. Lists can communicate only three logical relationships: sequence (first to last in time); priority (least to most important or vice versa); or simple membership in a set (these items relate to one another in some way, but the nature of that relationship remains unstated). And a list can show only one of those relationships at a time.¹⁷

¹⁷ Gordon Shaw, Robert Brown, Philip Bromiley, "Strategic Stories: How 3M is Rewriting Business Planning," *Harvard Business Review*, 76 (May-June, 1998), 44.

Shaw, Brown, and Bromiley found bullets leave "critical assumptions about how the business works unstated," and also displace narratives, an effective tool for thinking and for presentations. They describe, as we saw in the previous chapter on evidence corruption, the weakness of bullet outlines for thinking about causality, the fundamental idea behind strategic planning and, indeed, analytical thinking in general.

For scientists and engineers, a good way to help raise the quality of an analysis is to ask “What would Richard Feynman do?” The Feynman Principle can help with the presentation of scientific and engineering results. Feynman experienced the intense bullet outline style in his work on the first shuttle accident, the Challenger in 1986. He expressed his views clearly:

Then we learned about “bullets”—little black circles in front of phrases that were supposed to summarize things. There was one after another of these little goddamn bullets in our briefing books and on slides.¹⁸

¹⁸ Richard P. Feynman, “What Do You Care What Other People Think?” (New York, 1988), 126-127.

As analysis becomes more causal, multivariate, comparative, evidence-based, and resolution-intense, the more damaging the bullet list becomes. Scientists and engineers have communicated about complex matters for centuries without bullets and without PP. Richard Feynman wrote about much of physics—from classical mechanics to quantum electrodynamics—in 3 textbook volumes totalling 1,800 pages. These books use no bullets and only 2 levels of hierarchy, chapters and subheads within chapters:

front is an integral number of wavelengths. This difference can be seen to be $2d \sin \theta$, where d is the perpendicular distance between the planes. Thus the condition for coherent reflection is

$$2d \sin \theta = n\lambda \quad (n = 1, 2, \dots) \quad (38.9)$$

If, for example, the crystal is such that the atoms happen to lie on planes obeying condition (38.9) with $n = 1$, then there will be a strong reflection. If, on the other hand, there are other atoms of the same nature (equal in density) halfway between, then the intermediate planes will also scatter equally strongly and will interfere with the others and produce no effect. So d in (38.9) must refer to *adjacent* planes; we cannot take a plane five layers farther back and use this formula!

As a matter of interest, actual crystals are not usually as simple as a single kind of atom repeated in a certain way. Instead, if we make a two-dimensional analog, they are much like wallpaper, in which there is some kind of figure which repeats all over the wallpaper. By “figure” we mean, in the case of atoms, some arrangement—calcium and a carbon and three oxygens, etc., for calcium carbonate, and so on—which may involve a relatively large number of atoms. But whatever it is, the figure is repeated in a pattern. This basic figure is called a *unit cell*.

The basic pattern of repetition defines what we call the *lattice type*; the lattice type can be immediately determined by looking at the reflections and seeing what their symmetry is. In other words, where we find any reflections *at all* determines the lattice type, but in order to determine what is in each of the elements of the lattice one must take into account the *intensity* of the scattering at the various directions. *Which* directions scatter depends on the type of lattice, but *how strongly* each scatters is determined by what is inside each unit cell, and in that way the structure of crystals is worked out.

Two photographs of x-ray diffraction patterns are shown in Figs. 38-5 and 38-6; they illustrate scattering from rock salt and myoglobin, respectively.

Incidentally, an interesting thing happens if the spacings of the nearest planes are less than $\lambda/2$. In this case (38.9) has no solution for n . Thus if λ is bigger than twice the distance between adjacent planes then there is no side diffraction pattern, and the light—or whatever it is—will go right through the material without bouncing off or getting lost. So in the case of light, where λ is much bigger than the spacing, of course it does go through and there is no pattern of reflection from the planes of the crystal.

This fact also has an interesting consequence in the case of piles which make neutrons (these are obviously particles, for anybody’s money!). If we take these neutrons and let them into a long block of graphite, the neutrons diffuse and work their way along (Fig. 38-7). They diffuse because they are bounced by the atoms, but strictly, in the wave theory, they are bounced by the atoms because of diffraction from the crystal planes. It turns out that if we take a very long piece of graphite, the neutrons that come out the far end are all of long wavelength! In fact, if one plots the intensity as a function of wavelength, we get nothing except for wavelengths longer than a certain minimum (Fig. 38-8). In other words, we can get very slow neutrons that way. Only the slowest neutrons come through; they are not diffracted or scattered by the crystal planes of the graphite, but keep going right through like light through glass, and are not scattered out the sides. There are many other demonstrations of the reality of neutron waves and waves of other particles.

38-4 The size of an atom

We now consider another application of the uncertainty relation, Eq. (38.3). It must not be taken too seriously; the idea is right but the analysis is not very accurate. The idea has to do with the determination of the size of atoms, and the fact that, classically, the electrons would radiate light and spiral in until they settle down right on top of the nucleus. But that cannot be right quantum-mechanically because then we would know where each electron was and how fast it was moving.

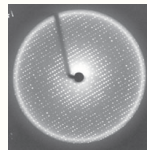


Figure 38-5

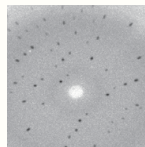


Figure 38-6

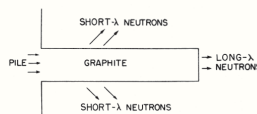


Fig. 38-7. Diffusion of pile neutrons through graphite block.

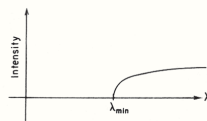


Fig. 38-8. Intensity of neutrons out of graphite rod as function of wavelength.

Page layout from Richard P. Feynman, Robert B. Leighton, and Matthew Sands, *The Feynman Lectures on Physics* (Reading, Massachusetts, 1963), volume 1, 38-5.

*The Gettysburg PowerPoint Presentation
by Peter Norvig*

The PP cognitive style is so distinctive and peculiar that presentations relying on standard ready-made templates sometimes appear as over-the-top parodies instead of the sad realities they are. Here is an intentional and ferocious parody: imagine Abraham Lincoln had used PowerPoint at Gettysburg. . . .

*Um, my name is Abraham Lincoln and, um,
I must now reboot*

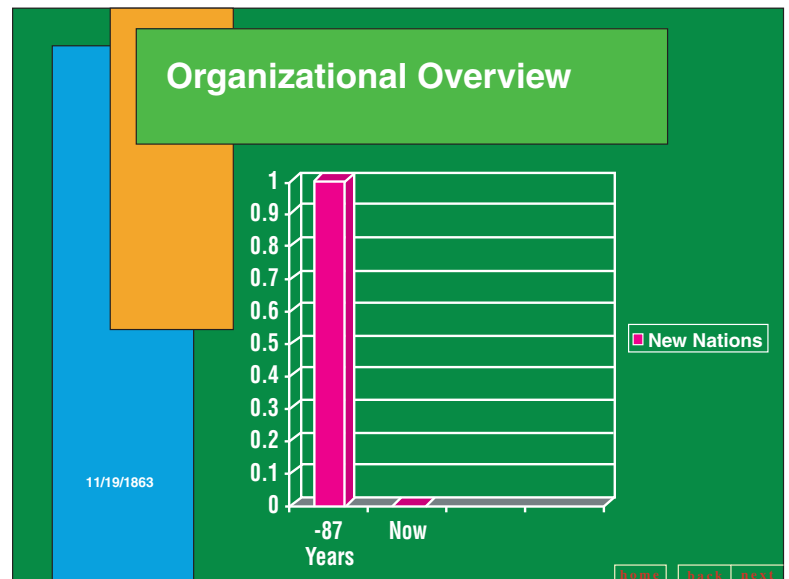
As we see in the Organizational Overview slide, four score and seven years ago our fathers brought forth on this continent a new nation, conceived in liberty and dedicated to the proposition that all men are created equal. Now we are engaged in a great civil war, testing whether that nation or any nation so conceived and so dedicated can long endure. Next slide please. We are met on a great battlefield of that war. We have come to dedicate a portion of that field as a final resting place for those who here gave their lives that that nation might live. It is altogether fitting and proper that we should do this. But in a larger sense, we cannot dedicate, we cannot consecrate, we cannot hallow this ground. The brave men, living and dead who struggled here have consecrated it far above our poor power to add or detract. Next slide please. The world will little note nor long remember what we say here, but it can never forget what they did here. It is for us the living rather to be dedicated here to the unfinished work which they who fought here have thus far so

11/19/1863

Gettysburg Cemetery Dedication

Abraham Lincoln

home back next



11/19/1863

Agenda

- Met on battlefield (great)
- Dedicate portion of field - fitting!
- Unfinished work (great tasks)

home back next

nobly advanced. It is rather for us to be here dedicated to the great task remaining before us— that from these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion, next slide please, that we here highly resolve that these dead shall not have died in vain, that this nation under God shall have a new birth of freedom, and that government of the people, by the people, for the people, shall not perish from the earth.

11/19/1863

Review of Key Objectives & Critical Success Factors

- What makes nation unique
 - Conceived in Liberty
 - Men are equal
- Shared vision
 - New birth of freedom
 - Gov't of/for/by the people

Home Back Next

This PowerPoint presentation was created by Peter Norvig; see www.norvig.com. The graph showing “-87 years” for Lincoln’s “four score and seven years ago” is brilliant. Norvig notes that other slides were quickly constructed by means of the PP AutoContent Wizard. Ian Parker described PowerPoint’s AutoContent Wizard as “a rare example of a product named in outright mockery of its target customers” (*The New Yorker*, May 28, 2001, 76).

11/19/1863

Not on Agenda!

- Dedicate
- Consecrate
- Hallow (in narrow sense)
- Add or detract
- Note or remember what we say

Home Back Next

11/19/1863

Summary

- New nation
- Civil War
- Dedicate field
- Dedicated to unfinished work
- New birth of freedom
- Government not perish

Home Back Next

PowerPoint and Statistical Evidence

To investigate the performance of PP for statistical data, let us consider an important and intriguing table of cancer survival rates relative to those without cancer for the same time period. Some 196 numbers and 57 words describe survival rates and their standard errors for 24 cancers:

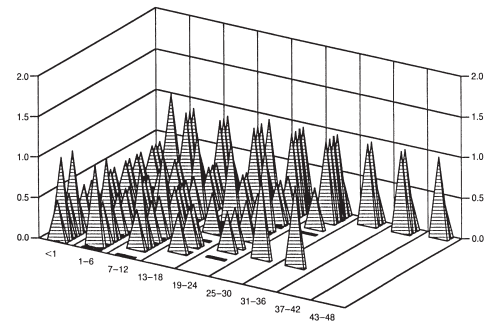
Estimates of relative survival rates, by cancer site¹⁹

	% survival rates and their standard errors							
	5 year		10 year		15 year		20 year	
Prostate	98.8	0.4	95.2	0.9	87.1	1.7	81.1	3.0
Thyroid	96.0	0.8	95.8	1.2	94.0	1.6	95.4	2.1
Testis	94.7	1.1	94.0	1.3	91.1	1.8	88.2	2.3
Melanomas	89.0	0.8	86.7	1.1	83.5	1.5	82.8	1.9
Breast	86.4	0.4	78.3	0.6	71.3	0.7	65.0	1.0
Hodgkin's disease	85.1	1.7	79.8	2.0	73.8	2.4	67.1	2.8
Corpus uteri, uterus	84.3	1.0	83.2	1.3	80.8	1.7	79.2	2.0
Urinary, bladder	82.1	1.0	76.2	1.4	70.3	1.9	67.9	2.4
Cervix, uteri	70.5	1.6	64.1	1.8	62.8	2.1	60.0	2.4
Larynx	68.8	2.1	56.7	2.5	45.8	2.8	37.8	3.1
Rectum	62.6	1.2	55.2	1.4	51.8	1.8	49.2	2.3
Kidney, renal pelvis	61.8	1.3	54.4	1.6	49.8	2.0	47.3	2.6
Colon	61.7	0.8	55.4	1.0	53.9	1.2	52.3	1.6
Non-Hodgkin's	57.8	1.0	46.3	1.2	38.3	1.4	34.3	1.7
Oral cavity, pharynx	56.7	1.3	44.2	1.4	37.5	1.6	33.0	1.8
Ovary	55.0	1.3	49.3	1.6	49.9	1.9	49.6	2.4
Leukemia	42.5	1.2	32.4	1.3	29.7	1.5	26.2	1.7
Brain, nervous system	32.0	1.4	29.2	1.5	27.6	1.6	26.1	1.9
Multiple myeloma	29.5	1.6	12.7	1.5	7.0	1.3	4.8	1.5
Stomach	23.8	1.3	19.4	1.4	19.0	1.7	14.9	1.9
Lung and bronchus	15.0	0.4	10.6	0.4	8.1	0.4	6.5	0.4
Esophagus	14.2	1.4	7.9	1.3	7.7	1.6	5.4	2.0
Liver, bile duct	7.5	1.1	5.8	1.2	6.3	1.5	7.6	2.0
Pancreas	4.0	0.5	3.0	1.5	2.7	0.6	2.7	0.8

¹⁹ Redesigned table based on Hermann Brenner, "Long-term survival rates of cancer patients achieved by the end of the 20th century: a period analysis," *The Lancet*, 360 (12 October 2002), 1131-1135. Brenner recalculates survival rates from data collected by the U.S. National Cancer Institute, 1973-1998, from the Surveillance, Epidemiology, and End Results Program.

Applying the PowerPoint templates for statistical graphics to this nice straightforward table yields the analytical disasters on the facing page. These PP default-designs cause the data to explode into 6 separate chaotic slides, consuming 2.9 times the area of the table. *Everything* is wrong with these smarmy, incoherent graphs: uncomparative, thin data-density, chartjunk, encoded legends, meaningless color, logotype branding, indifference to content and evidence. Chartjunk is a clear sign of statistical stupidity; use these designs in your presentation, and your audience will quickly and correctly conclude that you don't know much about data and evidence.²⁰ Poking a finger into the eye of thought, these data graphics would turn into a nasty travesty if used for

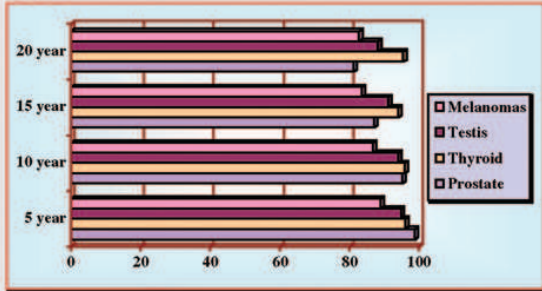
²⁰ PP-style chartjunk occasionally shows up in graphics of evidence in scientific journals. Below, the clutter half-conceals the thin data with some vibrating pyramids framed by an unintentional Necker illusion, as the 2 back planes optically flip to the front:



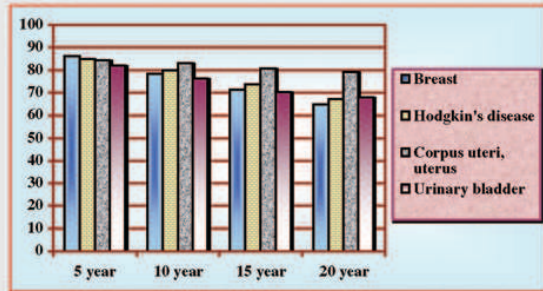
For such small data sets, usually a simple table will show the data more effectively than a graph, let alone a chartjunk graph. Source of graph: N. T. Kouchoukos, *et al.*, "Replacement of the Aortic Root with a Pulmonary Autograft in Children and Young Adults with Aortic-Valve Disease," *New England Journal of Medicine*, 330 (January 6, 1994), 4. On chartjunk, see Edward R. Tufte, *The Visual Display of Quantitative Information* (1983, 2001), chapter 5.



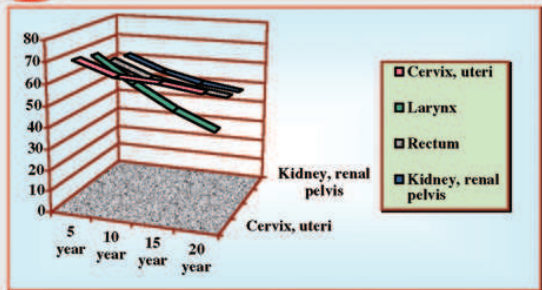
I. Cancer Survival Rates



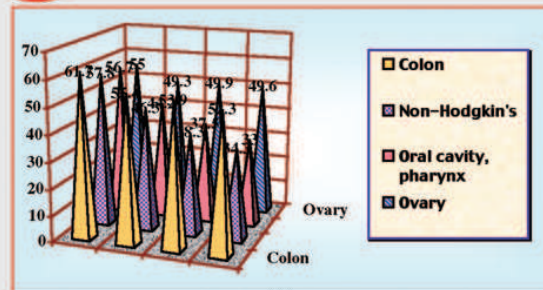
II. Cancer Survival Rates



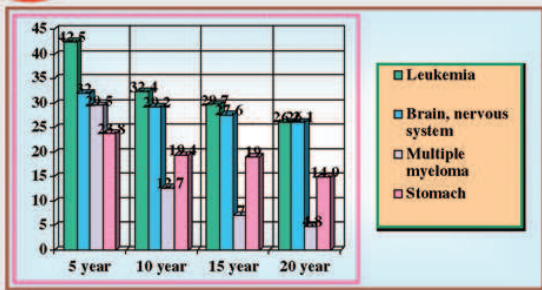
III. Cancer Survival Rates



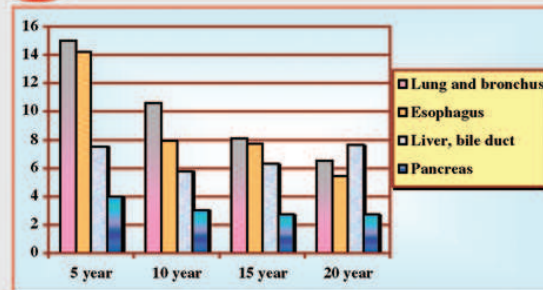
IV. Cancer Survival Rates



V. Cancer Survival Rates

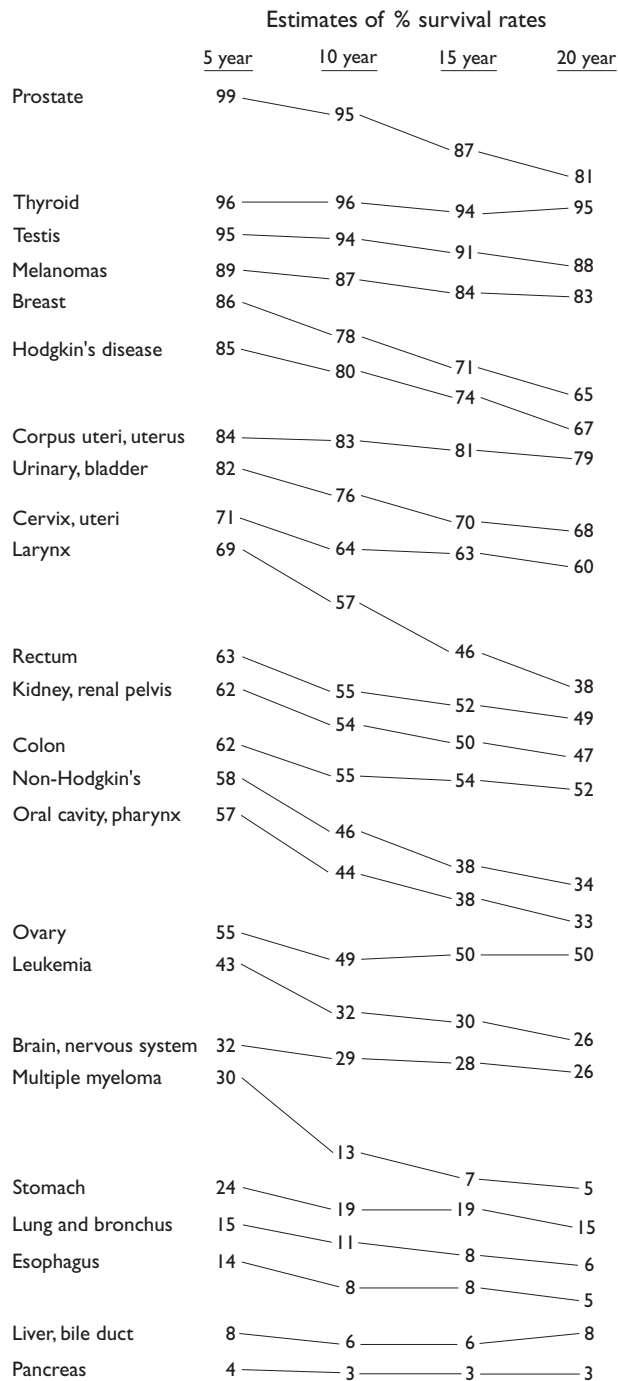


VI. Cancer Survival Rates



a serious purpose, such as cancer patients seeking to assess their survival chances. To deal with a product that messes up data with such systematic intensity must require an enormous insulation from statistical integrity and statistical reasoning by Microsoft PP executives and programmers, PP textbook writers, and presenters of such chartjunk.

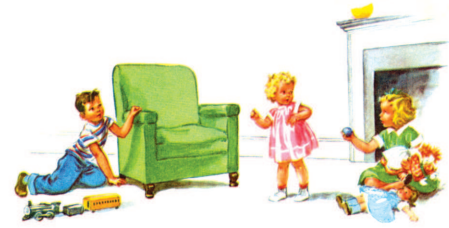
The best way to show the cancer data is the original table with its good comparative structure and reporting of standard errors. And PP default graphics are not the way to see the data. Our table-graphic, however, does give something of a *visual idea* of time-gradients for survival for each cancer. Like the original table, every visual element in the graphic shows data. Slideware displays, in contrast, usually devote a majority of their space to things other than data.



PowerPoint Stylesheets

THE PP cognitive style is propagated by the templates, textbooks, stylesheets, and complete pitches available for purchase. Some corporations and government agencies *require* employees to use designated PP Phluff and presentation logo-wear. With their strict generic formats, these designer stylesheets serve only to enforce the limitations of PowerPoint, compromising the presenter, the content, and, ultimately, the audience.

Here we see a witless PP pitch on how to make a witless PP pitch. Prepared at the Harvard School of Public Health by the “Instructional Computing Facility,” these templates are uninformed by the practices of scientific publication and the rich intellectual history of evidence and analysis in public health. The templates do, however, emulate the format of reading primers for 6 year-olds.



Jane said, “Here is a ball.
See this blue ball, Sally.
Do you want this ball?”

Sally said, “I want my ball.
My ball is yellow.
It is a big, pretty ball.”

Instructional Computing Facility

Guidelines for Preparing Slides

Instructional Computing
Facility

Harvard School of Public Health

Stylesheet-makers often seek to leave *their* name on *your* show; “branding,” as they say in the Marketing Department. In case you didn’t notice, this presentation is from the “Instructional Computing Facility.” But where are the names of the people responsible for this? No names appear on any of the 21 slides.

Instructional Computing Facility

Use the 6 X 6 rule:

6 lines of text
6 words per line

Harvard School of Public Health

This must be the Haiku Rule for formatting scientific lectures. At least we’re not limited to 17 syllables per slide. Above this slide, the rule can be seen in action—in a first-grade reading primer. The stylesheet typography, distinctly unscientific, uses a capital X instead of a multiplication sign.

Instructional Computing Facility

No More than One Topic per
Slide

What about them Sox hey?

Harvard School of Public Health

But this breaks up the evidence into arbitrary fragments. Why aren’t we seeing examples from actual scientific reports? What are the Sox (a rather parochial reference) doing here? The inept PP typography persists: strange over-active indents, oddly chosen initial caps, typographic orphans on 3 of 4 slides.

Instructional Computing Facility

Outline Formats are Easier to
Follow

Harvard School of Public Health

Why is this relevant to scientific presentations? Are there other principles than ease of following? Didn’t the *Harvard Business Review* article indicate that bullet outlines corrupt thought? Text, imaging, and data for scientific presentations should be at the level of scientific journals, much *higher* resolution than speech.

Instructional Computing Facility

Use Simple Tables to Present Numbers

	Use Tables	For Your Numbers	But Not too Many
This row	10	90	100
This row	0.6	0.4	1
This row	1	2	3
That row	1	2	3

Try not to make footnotes too small

Harvard School of Public Health

The stylesheet goes on to victimize statistical data, the fundamental evidence of public health. The table shows 12 numbers which is lousy for science, sports, weather, or financial data but standard for PowerPoint.²¹ Table design is a complex and subtle matter in typographic work, but there is nothing thoughtful about design here. The unsourced numbers are not properly aligned, the row and column labels are awful, the units of measurement not given. This stylesheet of pseudoscience displays a flippant smirky attitude toward evidence. That attitude — *what counts are power and pitches, not truth and evidence* — also lurks within PowerPoint.

Consider now a real table. Bringing scientific methods to medical and demographic evidence, John Graunt's *Bills of Mortality* (1662) is the foundation work of public health. Graunt calculated the first tables of life expectancy, compared different causes of death, and even discussed defects in the evidence. His renowned "Table of Casualties" (at right) shows 1,855 different counts of death from 1629 to 1659. How fortunate that Graunt did not have PowerPoint and the assistance of the Harvard School of Public Health Instructional Computing Facility. Their silly guidelines above suggest the construction of 155 separate PowerPoint slides to show the data in Graunt's original table!

For tables, the analytical idea is to make comparisons. The number of possible pairwise comparisons in a table increases as the square of the number of cells.²² In Graunt's table, 1,719,585 pairwise comparisons, of varying relevance to be sure, are within the eyespan of the inquiring mind. In contrast, the 155 tiny tables on 155 PP slides would offer only 10,230 pairwise comparisons, about 6 in 1,000 of those available in Graunt's original table. These PP tables would also block all sorts of interesting comparisons, such as time patterns over many years. What Graunt needs to do for his presentation at Harvard is simply to provide printed copies of his original table to everyone in the audience.

²¹ Some 39 tables appear in our collection of 28 PP textbooks. These tables show an average (median) of 12 numbers each, which approaches the *Pravda* level. In contrast, sports and financial pages in newspapers routinely present tables with hundreds, even thousands of numbers. Below, we see a conventional weather table from a newspaper. The Harvard School of Public Health PP guidelines inform presenters that this data set will require 31 PP slides:

Africa	Yesterday	Today	Tomorrow
Algiers	82/ 66 0.55	85/ 60 S	85/ 61 S
Cairo	99/ 70 0	101/ 76 S	96/ 76 S
Cape Town	64/ 54 0.16	63/ 49 PC	60/ 50 Sh
Dakar	87/ 77 0.75	86/ 81 PC	85/ 81 PC
Johannesburg	69/ 42 0	73/ 42 S	71/ 47 S
Nairobi	75/ 55 0	78/ 56 PC	78/ 56 PC
Tunis	80/ 69 -	87/ 73 PC	85/ 71 PC
Asia/Pacific	Yesterday	Today	Tomorrow
Auckland	59/ 45 0.12	58/ 44 Sh	58/ 44 Sh
Bangkok	91/ 82 0	91/ 79 Sh	91/ 77 Sh
Beijing	85/ 57 0	84/ 60 S	78/ 65 PC
Bombay	88/ 75 0.28	87/ 77 T	88/ 78 T
Damascus	96/ 55 0	98/ 59 S	96/ 62 S
Hong Kong	91/ 77 0	88/ 81 PC	92/ 78 PC
Jakarta	89/ 77 0	90/ 77 PC	89/ 77 PC
Jerusalem	87/ 64 0	88/ 66 S	88/ 69 S
Karachi	86/ 80 0	92/ 78 PC	92/ 79 S
Manila	86/ 75 -	84/ 75 R	87/ 78 R
New Delhi	89/ 80 Tr	88/ 76 Sh	92/ 76 Sh
Riyadh	98/ 69 0	102/ 74 S	101/ 75 S
Seoul	78/ 64 2.09	83/ 65 PC	77/ 66 R
Shanghai	75/ 69 0.06	86/ 76 Sh	86/ 73 PC
Singapore	87/ 78 Tr	89/ 76 R	89/ 78 Sh
Sydney	68/ 53 0	71/ 51 PC	71/ 48 PC
Taipei	84/ 77 2.28	87/ 73 PC	88/ 72 PC
Tehran	93/ 73 0	87/ 73 S	87/ 73 S
Tokyo	89/ 77 0	91/ 79 Sh	83/ 80 Sh
Europe	Yesterday	Today	Tomorrow
Amsterdam	56/ 50 0.39	66/ 51 PC	64/ 52 Sh
Athens	87/ 75 0	90/ 75 S	88/ 71 S
Berlin	64/ 55 0.31	61/ 49 R	68/ 52 PC
Brussels	62/ 54 Tr	66/ 53 PC	65/ 52 Sh
Budapest	72/ 59 0	75/ 55 S	67/ 53 Sh
Copenhagen	59/ 51 0.08	63/ 51 Sh	63/ 52 PC
Dublin	66/ 54 0.12	66/ 55 Sh	63/ 47 PC
Edinburgh	63/ 46 0.02	63/ 46 R	64/ 48 PC
Frankfurt	65/ 54 0.01	65/ 54 Sh	66/ 50 PC
Geneva	69/ 57 0.04	64/ 56 Sh	65/ 50 PC
Helsinki	63/ 45 0	62/ 46 PC	63/ 45 PC
Istanbul	84/ 60 0.01	79/ 69 Sh	78/ 67 S
Kiev	66/ 46 0	64/ 47 S	64/ 46 S
Lisbon	84/ 62 0	91/ 65 S	90/ 67 S
London	71/ 53 0.08	66/ 53 Sh	69/ 55 PC
Madrid	86/ 46 0	87/ 55 S	87/ 57 S
Moscow	55/ 41 0	64/ 40 S	62/ 44 S
Nice	78/ 62 0.01	78/ 65 S	78/ 63 S
Oslo	62/ 48 0	57/ 47 PC	59/ 45 PC
Paris	68/ 57 0	69/ 56 PC	68/ 57 PC
Prague	64/ 55 0.04	56/ 49 T	63/ 49 Sh
Rome	75/ 62 -	79/ 61 S	76/ 60 Sh
St. Petersburg	59/ 39 0	66/ 46 S	65/ 47 PC
Stockholm	64/ 46 0	61/ 49 PC	63/ 45 PC
Vienna	64/ 59 0.16	65/ 53 PC	66/ 52 Sh
Warsaw	69/ 46 0	62/ 51 Sh	65/ 49 PC

²² A table with n cells yields $n(n - 1)/2$ pairwise comparisons of cell entries.

John Graunt, *National and Political Observations mentioned in a following index, and made upon the Bills of Mortality. With reference to the Government, Religion, Trade, Growth, Ayre, Diseases, and the several Changes of the said City* (London, 1662); "The Table of Casualties" follows folio 74.

PP Slide Formats for Paper Reports and Computer Screens Are Ridiculous and Lazy

In addition to accompanying a talk, PP slides are printed out on paper, attached to emails, posted on the internet. Unfortunately, PP slides on paper and computer screens *replicate and intensify* all the problems of the PP cognitive style. Such slides extend the reach of PP's proprietary closed-document format since PP capabilities are necessary to see the slides. This short-run convenience to presenters and long-run benefit to Microsoft comes at an enormous cost to the content and the audience.

As those who have disconsolately flipped through pages and pages of printed-out PP slide decks already know, such reports are physically thick and intellectually thin. Recall that the NASA Return to Flight Task Group observed a massive thinness in the PP closure reports. The resolution of printed-out slide decks is remarkably low, approaching dementia. This data table compares the information in one image-equivalent for books (one page), for the internet (one screen), and for PP (one slide). A single page in the *Physicians' Desk Reference* shows 54 typical PP slide-equivalents of information, and the whole very thick book equals a deck of 181,000 slides. A single page of an Elmore Leonard novel equals 13 typical PP slides. Nonfiction best-sellers show information at densities 10 to 50 times those of printed-out PP decks.

People see, read, and think all the time at intensities vastly greater than those presented in printed PP slides. Instead of showing a long sequence of tiny information-fragments on slides, and instead of dumping those slides onto paper, report makers should have the courtesy to write a real report (which might also be handed out at a meeting) and address their readers as serious people. PP templates are a lazy and ridiculous way to format printed reports.

PP slides also format information on computer screens. Presenters post their slides; then readers, if any, march through one slide after another on the computer screen. Popular news sites on the internet show 10 to 15 times more information on a computer screen than a typical PP slide on a computer screen. The shuttle Columbia reports prepared by Boeing, sent by email in PP format to be viewed on computer screens, were running at information densities of 20% of major news sites on the internet, as the table shows.

The PP slide format has the worst signal/noise ratio of any known method of communication on paper or computer screen. Extending PowerPoint to embrace paper and internet screens pollutes those display methods.

CHARACTER COUNTS AND DENSITY PER PAGE-IMAGE

	CHARACTERS PER PAGE	DENSITY: CHARACTERS/IN ²
BEST SELLING BOOKS		
<i>Physicians' Desk Reference</i>	13,600	168
<i>Your Income Tax</i>	10,400	118
<i>World Almanac</i>	9,800	232
<i>Joy of Cooking</i>	5,700	108
<i>The Merck Manual</i>	4,700	117
<i>Guinness Book of World Records</i>	4,600	162
<i>Consumer Reports Buying Guide</i>	3,900	112
<i>How to Cook Everything</i>	3,900	53
<i>Maximum Bob</i> (Elmore Leonard)	3,100	115
<i>Baby and Child Care</i>	2,500	95
NEWS SITES ON THE INTERNET		
Google News	4,100	44
New York Times	4,100	43
People's Daily (China)	4,100	43
Pravda	4,100	43
Los Angeles Times	4,000	42
BBC News	3,400	36
CNN	3,300	35
Yahoo	3,200	34
Time	2,700	28
MSNBC	2,400	26
POWERPOINT SLIDE FORMAT USED ON PAPER OR COMPUTER SCREEN		
Columbia reports by Boeing	630	7
1,460 text slides in 189 PP reports	250	3
654 text slides in 28 PP textbooks	98	1
Content-free slides	0	0

Competitive Analysis of Presentation Tools

OUR comparisons of various presentation tools in action indicate that PowerPoint is intellectually outperformed by alternative tools. For the 10 case studies and 32 control samples, PP flunks the comparative tests, except for beating out *Pravda* in the statistical graphics competition.

Some of these comparisons are for *the same users with the same content*. Matched comparisons control for selection effects, such as the entertaining hypothesis that PP is a stupidity magnet, differentially attracting inept presenters with lightweight content (and thereby making PP look bad). Our evidence helps isolate PP effects, independent of user or content. Such comparisons—*Consumer Reports* style—provide a competitive analysis of presentation tools. In these tests, PP's poor performance cannot be blamed on its users. For example, in the shuttle investigations, given that the presenters are NASA engineers and the content is rocket science, which then is the better presentation method, PP or technical reports?

The scope of our evidence is limited. Nearly all the evidence is drawn from *serious presentations*, with explanations to understand, evidence to evaluate, problems to solve, decisions to make, and, in several examples, lives to save. It is hard to know how many presentations are serious. Perhaps 25% to 75%, depending very much upon the substantive field.

What Are the Causes of Visual Presentations?

AN important but complex issue in evaluating visual presentations, including PowerPoint, is *what are the causes of a presentation?* What are the contributions of content quality, presenter skills, presentation methods, cognitive styles, and prevailing standards of integrity? To begin with, reasonably certain answers are that the causal structure is multivariate, that causes tend to interact and are not independent of one another, and that improvements will result from working on all factors.

George Orwell's classic essay "Politics and the English Language" gets right the interplay between quality of thought and cognitive style of presentation: "The English language becomes ugly and inaccurate because our thoughts are foolish, but the slovenliness of our language makes it easier for us to have foolish thoughts." Imagine Orwell writing about PP: "PowerPoint becomes ugly and inaccurate because our thoughts are foolish, but the slovenliness of PowerPoint makes it easier for us to have foolish thoughts." The PP cognitive style is familiar to readers of Orwell's remarkable and prescient novel *1984*.

WAR IS PEACE

**WAR IS PEACE
FREEDOM IS SLAVERY**

**WAR IS PEACE
FREEDOM IS SLAVERY
IGNORANCE IS STRENGTH**

Or consider the NASA presentations. What are the causes of the dreaded Engineering by PowerPoint? Engineers incapable of communicating by means of standard technical reports? Lack of intellectual rigor? Designer guidelines and bureaucratic norms that insist on PP for all presentations, regardless of content? The cognitive style of PowerPoint? A bureaucracy infected throughout by the pitch culture? The PowerPoint monopoly and the consequent lack of innovative and high-quality software for technical communication? A Conway's Law interaction of causes? Some or all of these factors? In what proportion?

Sorting all this out is not possible. Nonetheless, under most reasonable allocations of causal responsibility, the practical advice remains the same: To make smarter presentations, try smarter tools. Technical reports are smarter than PowerPoint. Sentences are smarter than the grunts of bullet points. PP templates for statistical graphics and data tables are hopeless.

ART historians reason about the causes of visual presentations. What can we learn from their work? To explain artistic productions, art historians make use of 4 grand explanatory variables: (1) differences in styles in art, (2) differences in artists working within a given style, (3) interplay among artists and styles, and (4) sources of new styles.

The prevailing *style* of a particular place and period deeply affects the character of art work. Art history textbooks are written as narratives of distinctive, clearly identifiable styles: Prehistoric, Egyptian, Near Eastern, Classical, Byzantine, Islamic, Baroque, Renaissance, Far Eastern, African, Romanticism, Impressionism, Cubism, and many other distinct styles. In the long history of representational art, the represented objects did not change all that much, nor did artists' retinal images of those objects. The big changes in art resulted from changes in style. Style matters.

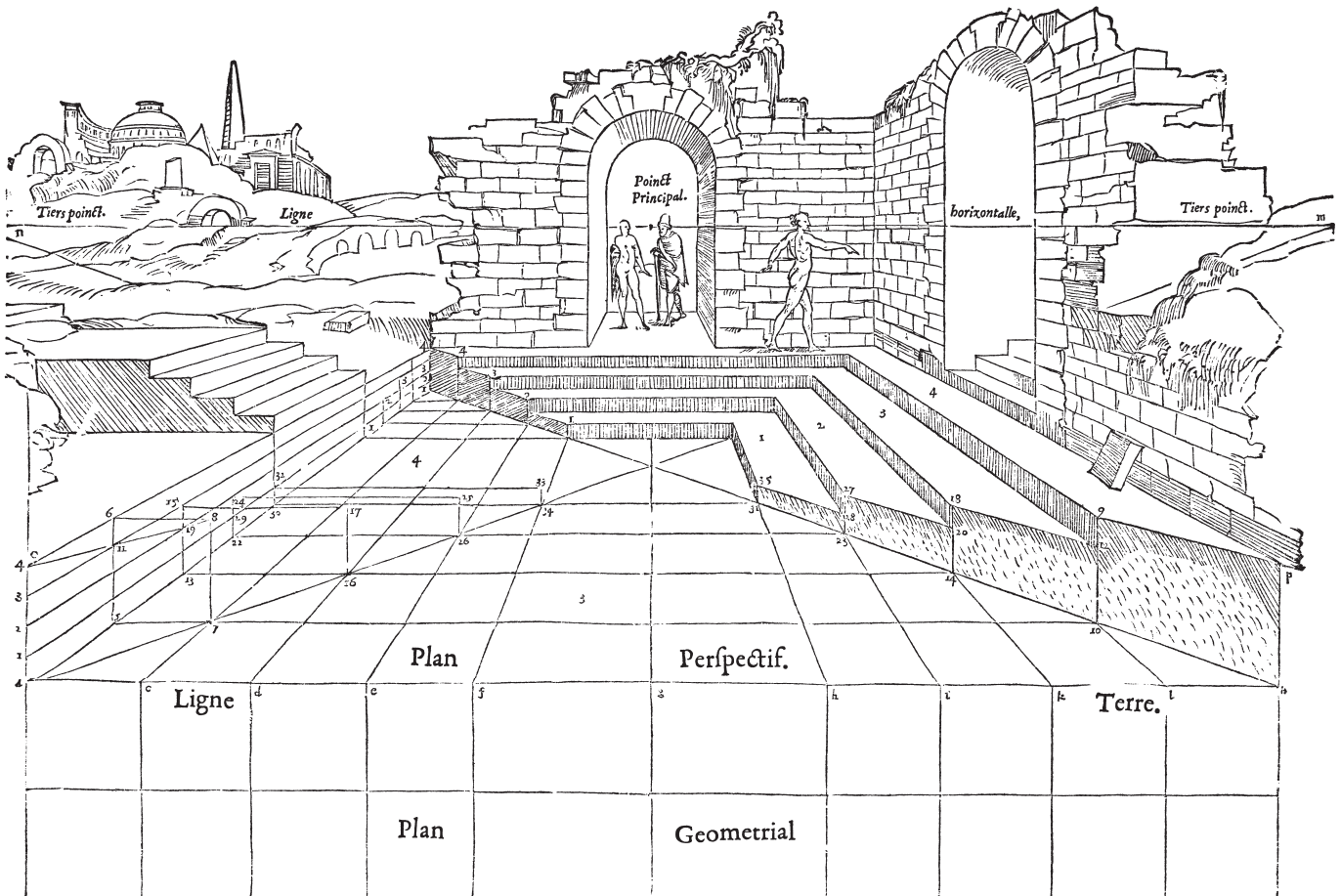
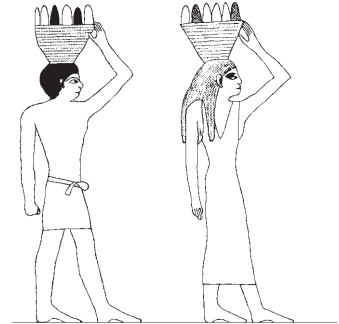
Those caught up *within a single style* of visual production, however, must necessarily explain differences in quality by reference to the skills and character of particular presenters, for style is a given. This is the method of the standard defense of PowerPoint, a defense that mobilizes the second grand explanatory variable, presenter variability, as the determinant of visual productions. Lousy presentations are said to be *the fault of inept PP users, not the fault of PP*. Blame the user, not the cognitive style of the presentation tool, not the PP pitch culture.

That is sometimes the case, but causal responsibility for presentations is more complicated than that. Other explanatory variables of visual productions—cognitive style and quality of the presentation tools, user-style interactions, context, character of the content—must be taken into account. Thus Orwell's Principle, for example, sensibly avoids mono-causal explanations: "The English language becomes ugly and inaccurate because our thoughts are foolish, but the slovenliness of our language makes it easier for us to have foolish thoughts." And so our comparisons

of the PP cognitive style with other tools; thus our analysis of the PP metaphors of marketing and hierarchy at work and play in bureaucracies.

What about modest incremental reforms in the cognitive style of PowerPoint? There are inherent problems in PP, and also the record is not promising. Throughout many versions of PP, the intellectual level and analytical quality has rarely improved. New releases feature more elaborated PP Phluff and therapeutic measures for troubled presenters. These self-parodying elaborations make each new release *different* from the previous version—but not smarter. PP competes largely with itself: there are few incentives for meaningful change in a monopoly product with an 86% gross profit margin (as reported in antitrust proceedings). In a competitive market, producers improve and diversify products; monopolies have the luxury of blaming consumers for poor performances. It is scandalous that there is no coherent software for serious presentations.

A better cognitive style for presentations is needed, a style that respects, encourages, and cooperates with evidence and thought. PowerPoint is like being trapped in the style of early Egyptian flatland cartoons rather than using the more effective tools of Renaissance visual representation.



Jean Cousin, *Livre de perspective* (Paris, 1560), I iij.

Improving Presentations

At a minimum, we should choose presentation tools that *do no harm* to content. Yet PowerPoint promotes a cognitive style that disrupts and trivializes evidence. PP presentations too often resemble a school play: very loud, very slow, and very simple. Since 10^{10} to 10^{11} PP slides are produced yearly, that is a lot of harm to communication with colleagues.

PowerPoint is a competent slide manager, but a Projector Operating System should not impose Microsoft's cognitive style on our presentations. PP has some occasionally competent low-end design tools and way too many Phluff tools. PP might help show a few talking points at informal meetings, but instead why not simply print out an agenda for everyone?

For serious presentations, replace PP with word-processing or page-layout software. Making this transition in large organizations requires a straightforward executive order: *From now on your presentation software is Microsoft Word, not PowerPoint. Get used to it.*

Someday there will be a good technical reporting tool. Focused on evidence analysis and display, this tool should combine a variety of page and screen layout templates (based on formats for serious news reports, an article in *Nature*, Feynman's physics textbook, and so on); publication-quality statistical graphics and tables; scientific notation and typography; graphics tools for placing annotated measurement scales in images; spellchecking for technical terms; *within-document* editing of words, tables, graphics, and images; *open-document* non-proprietary formats; fast color printing for large paper; and a slide manager for talks.

At a talk, paper handouts of a technical report effectively show text, data graphics, images. Printed materials bring information transfer rates in presentations up to that of everyday material in newspaper sports and financial pages, books, and internet news sites. An excellent paper size for presentation handouts is A3, 30 by 42 cm or about 11 by 17 inches, folded in half to make 4 pages. That one piece of paper, the 4-pager, can show images with 1,200 dpi resolution, up to 60,000 characters of words and numbers, detailed tables worthy of the sports pages, or 1,000 sparkline statistical graphics showing 500,000 numbers. *That one piece of paper shows the content-equivalent of 50 to 250 typical PP slides.* Thoughtful handouts at your talk demonstrate to the audience that you are responsible and seek to leave permanent traces and have consequences. Preparing a technical report requires deeper intellectual work than simply compiling a list of bullets on slides. Writing sentences forces presenters to be smarter. And presentations based on sentences make consumers smarter as well.

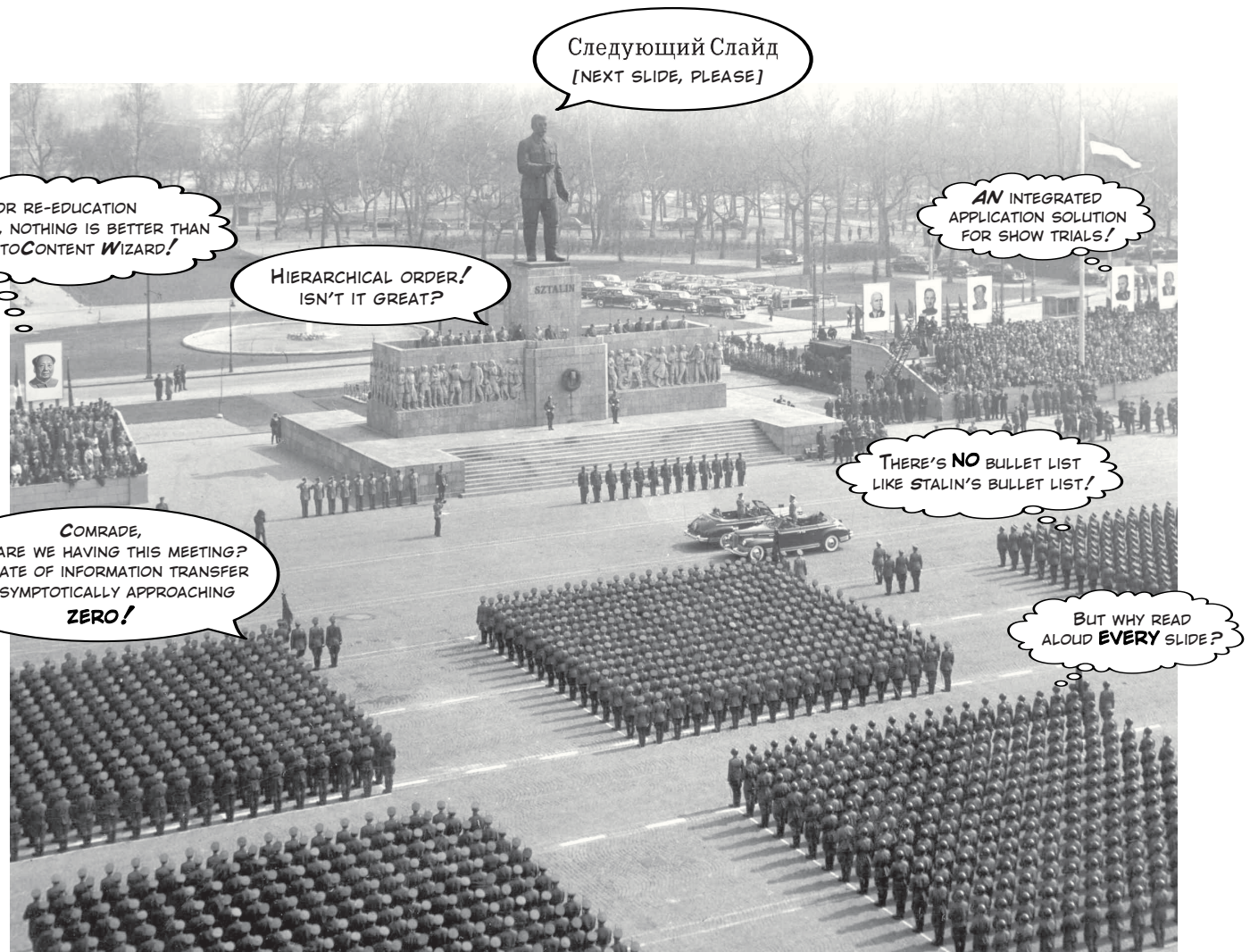
Serious presentations might well begin with a concise briefing paper or technical report (the 4-pager) that everyone reads (people can read 3 times faster than presenters can talk). Following the reading period, the presenter might provide a guided analysis of the briefing paper and then encourage and perhaps lead a discussion of the material at hand.

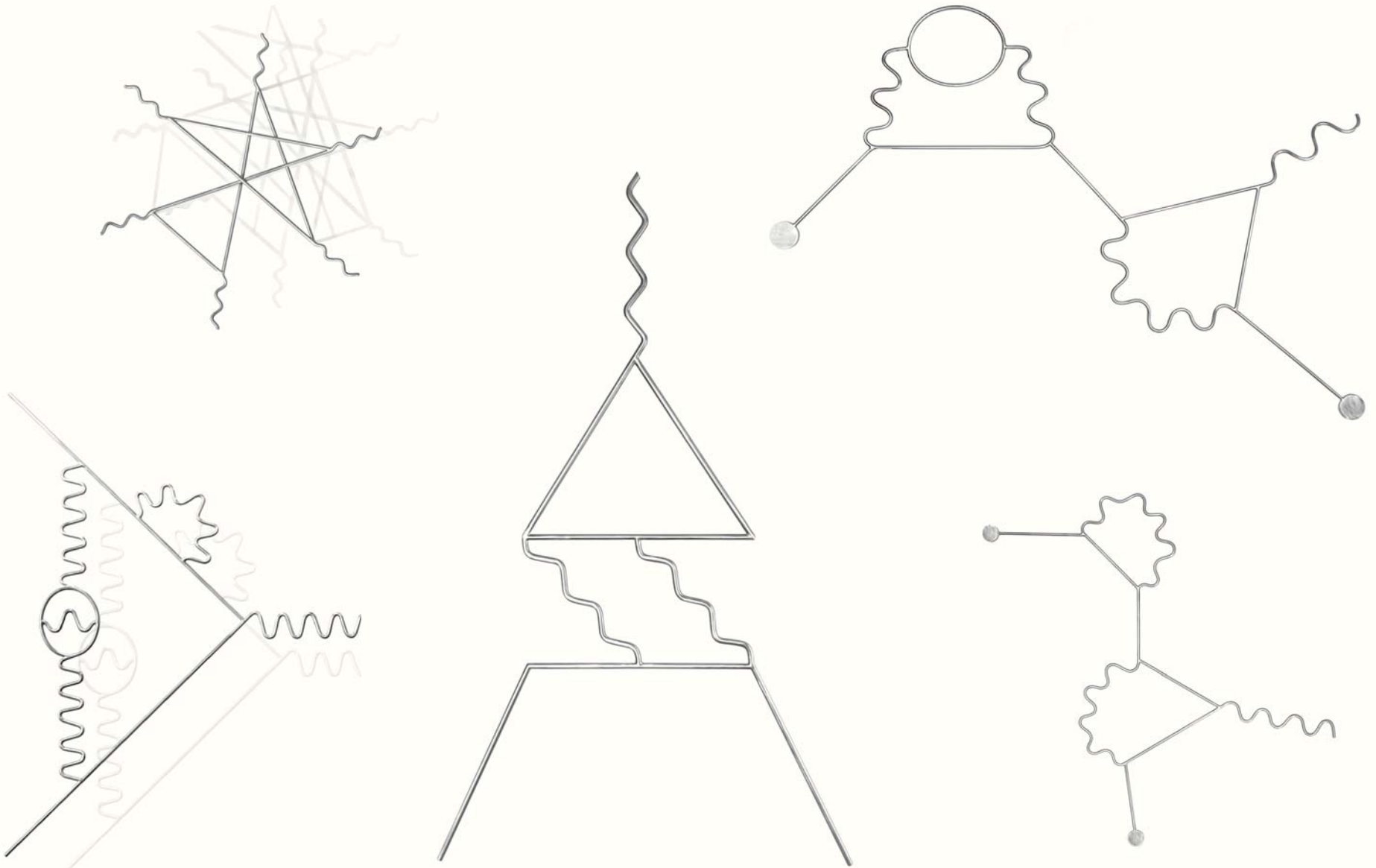
Consuming Presentations

OUR evidence concerning PP's performance is relevant only to serious presentations, where the audience needs (1) to understand something, (2) to assess the credibility of the presenter. For non-serious pitches and meetings, the PP cognitive style may not matter all that much. Rather than providing information, *PowerPoint allows speakers to pretend that they are giving a real talk, and audiences to pretend that they are listening.* This prankish conspiracy against evidence and thought should provoke the question, *Why are we having this meeting?*

Consumers of presentations might well be skeptical of speakers who rely on PowerPoint's cognitive style. It is possible that these speakers are not evidence-oriented, and are serving up some PP Phluff to mask their lousy content, just as this massive tendentious pedestal in Budapest once served up Stalin-cult propaganda to orderly followers feigning attention.

Military parade, Stalin Square, Budapest, April 4, 1956. Photograph ©Associated Press.

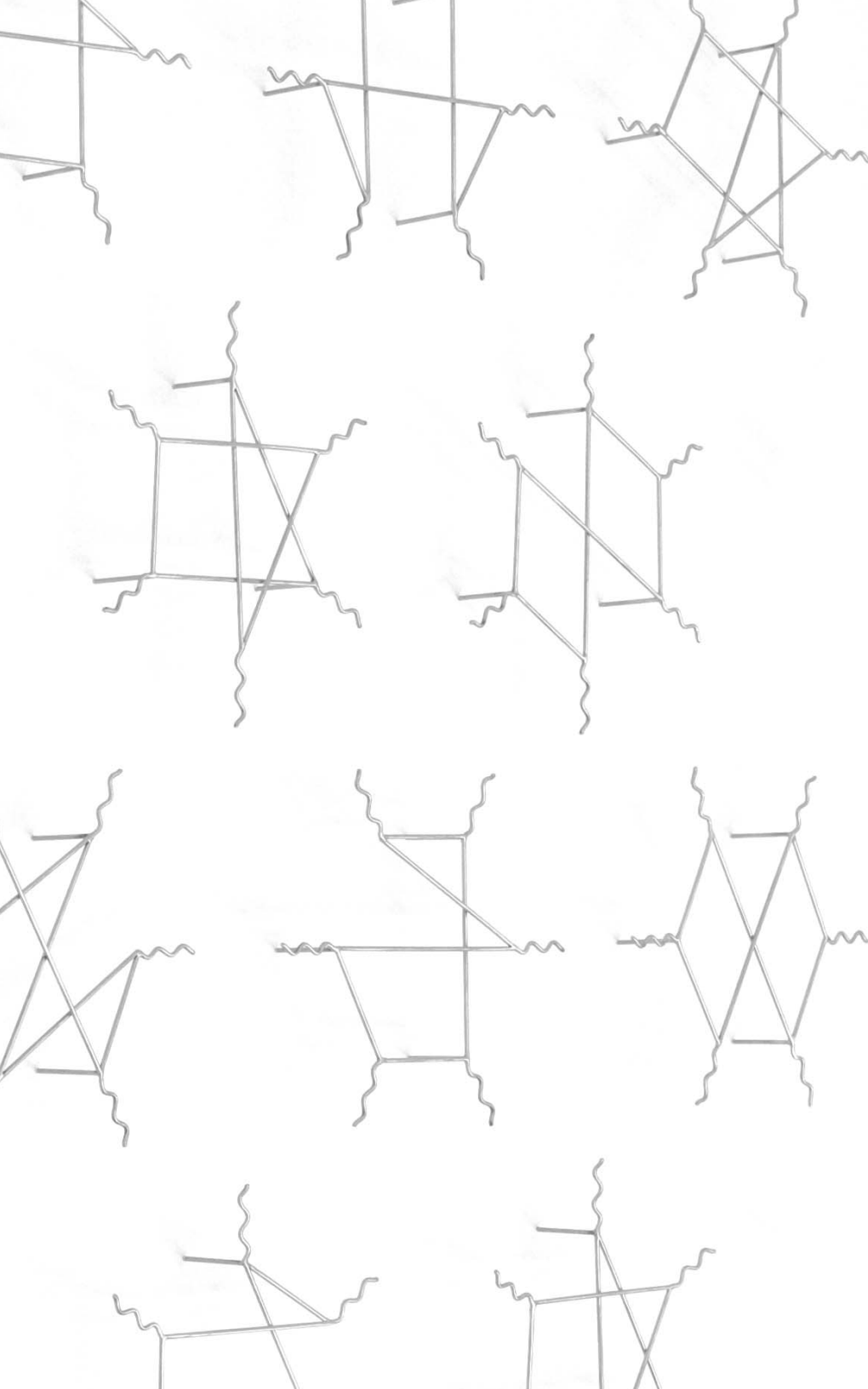




EDWARD TUFTE ALL POSSIBLE PHOTONS
THE CONCEPTUAL AND COGNITIVE ART OF FEYNMAN DIAGRAMS

ET MODERN 547 WEST 20TH STREET NEW YORK





Edward Tufte's wall-mounted sculptures, *All Possible Photons*, generate an enormous multiplicity of three-dimensional optical experiences of line, light, airspace, color, shadow, form.

Made from stainless steel and air, the artworks grow out of Richard Feynman's famous diagrams describing Nature's subatomic behavior. Feynman diagrams depict the space-time patterns of particles and waves of quantum electrodynamics. These mathematically derived and empirically verified visualizations represent the space-time paths taken by all subatomic particles in the universe.

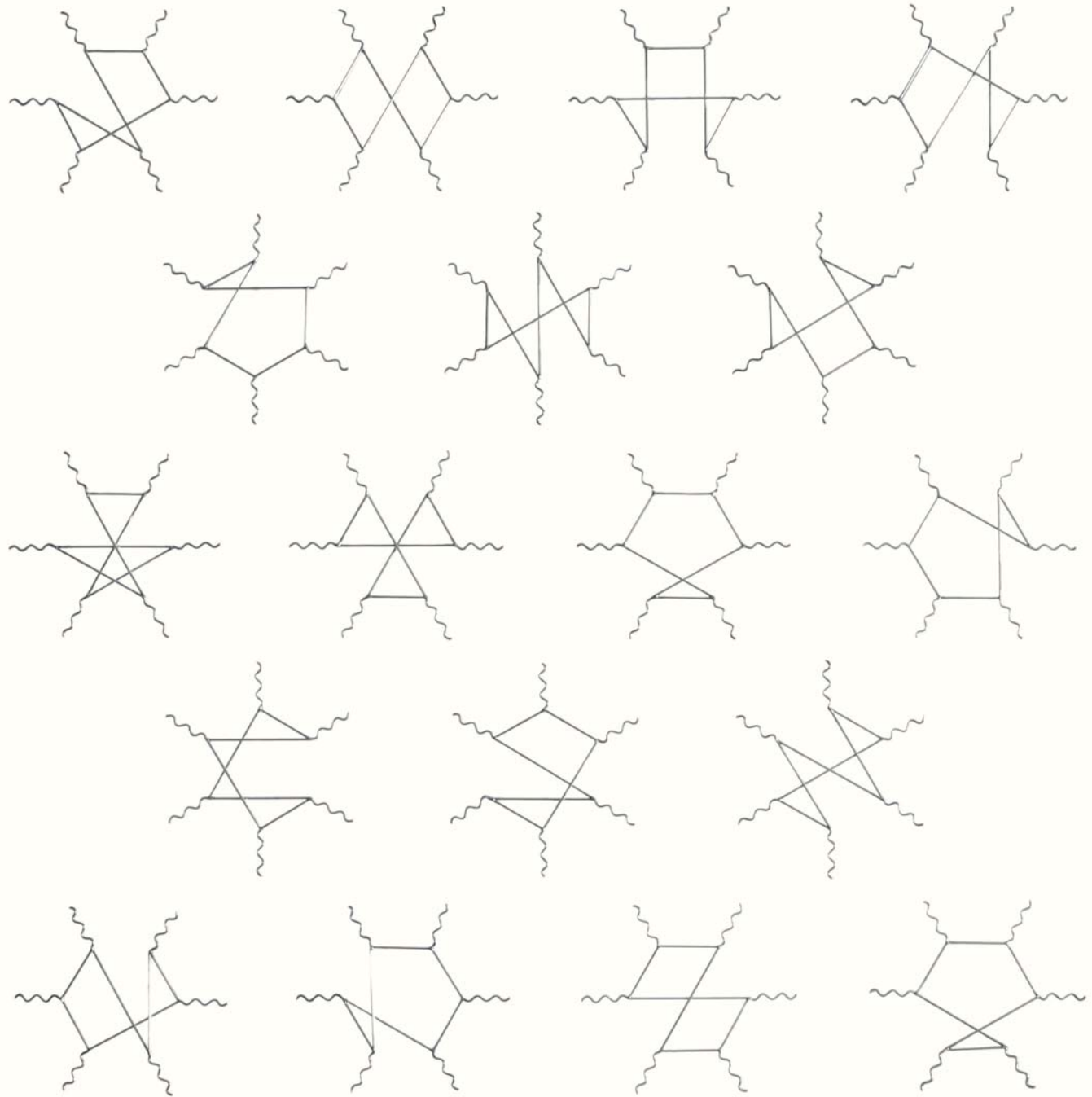
The resulting conceptual and cognitive art is both beautiful and true. Along with their art, the stainless steel elements of *All Possible Photons* actually *represent something*: the precise activities of Nature at her highest resolution.

Gathered together, as in the 120 diagrams showing all possible space-time paths of 6-photon scattering, the stainless steel lines (and their variable shadow, airspace, light, color, form) reveal the endless complexities that result from multiplying and varying fundamental elements.

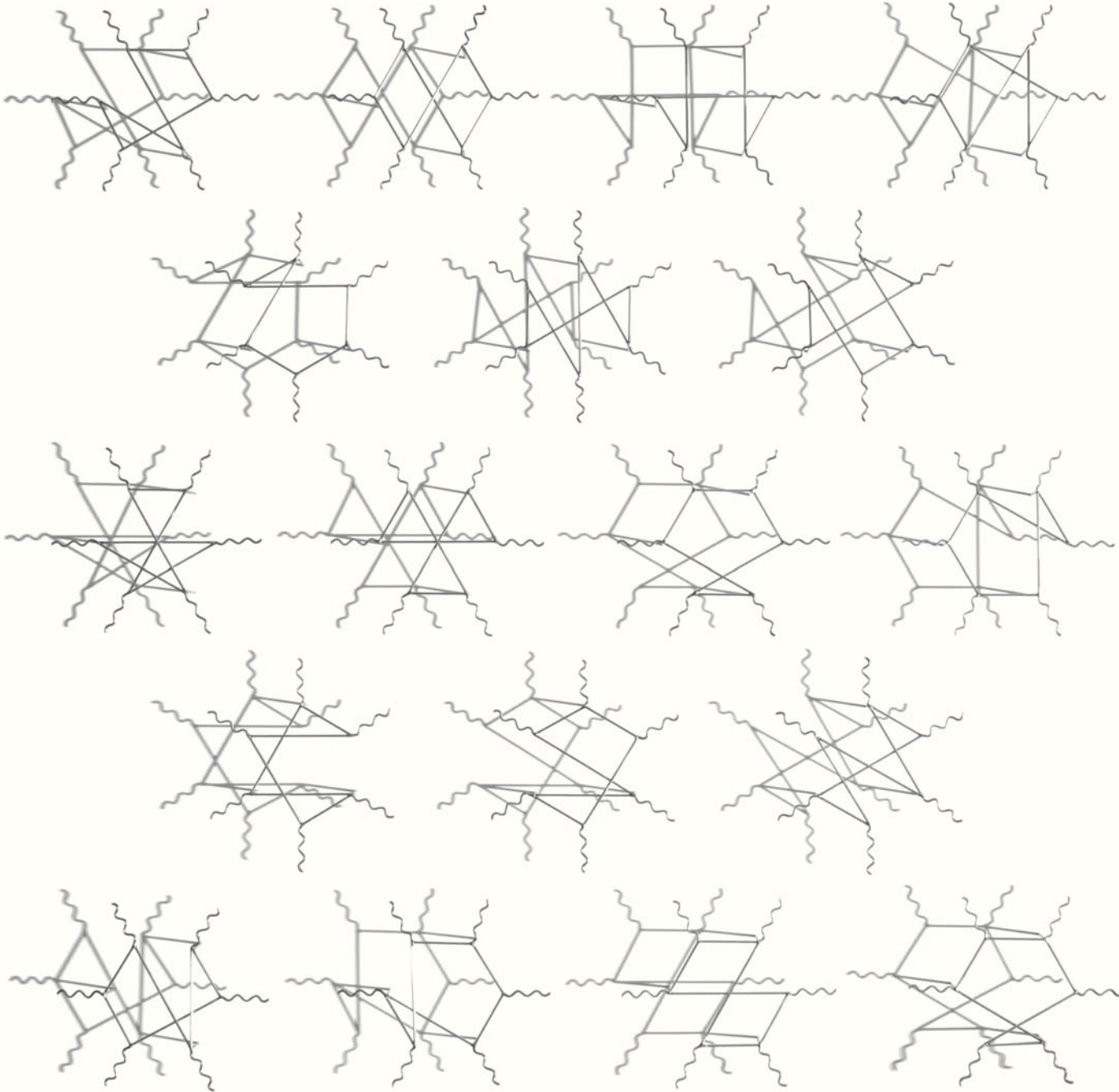
"How beautiful it was then," writes Italo Calvino about a time of radiant clarity in cosmic prehistory, "through that void, to draw lines and parabolas, pick out the precise point, the intersection between space and time when the event would spring forth, undeniable in the prominence of its glow."

ET *All Possible 6-Photon Scattering (120 Space-Time Feynman Diagrams)* 2012
stainless steel 17.5 x 7.3 x .2 feet or 5.3 x 2.3 x .1 meters (detail)

Without shadow light, the artwork reads as precise lines on a flat surface. But since all the 18 elements are supported off the wall, they cast shadows. (A good definition of sculpture is *artwork that casts shadows*.) These shadows are shown on the adjacent page.



*18 Space-Time Feynman Diagrams
of 6-Photon Scattering 2012*
stainless steel 4.1 x 4.2 x .2 feet
or 1.3 x 1.3 x .1 meters



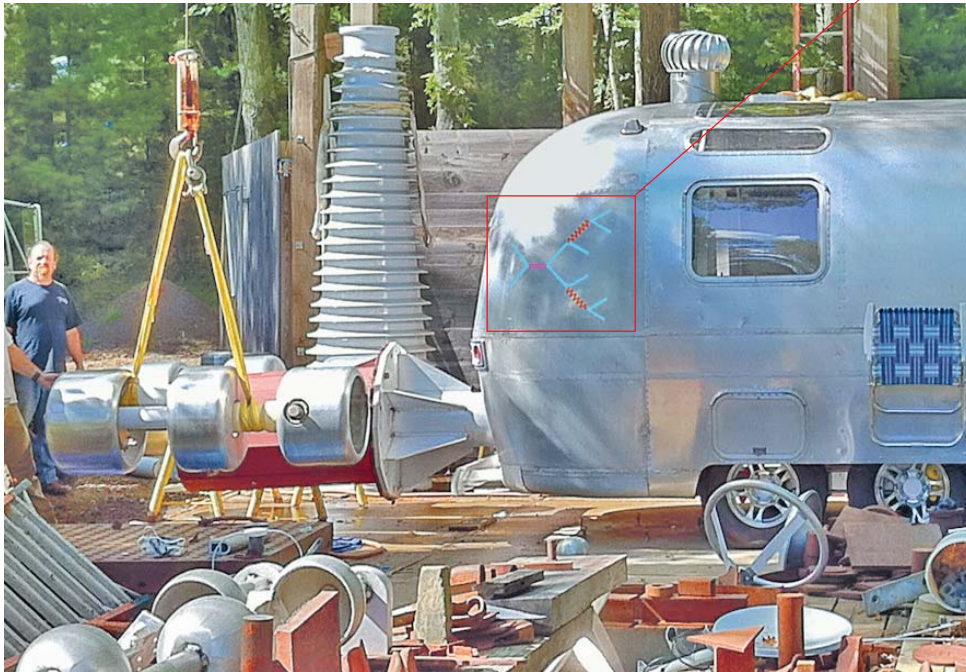
With light, the same artwork yields a luscious complexity: flat shadows interacting with 3D stainless steel lines to create Escher-impossible apparent objects in 3-space.

And when the light shifts in color, intensity, and angle of incidence, so the reflections from the steel lines and their shadows continually respond, vary, shift. These diverse optical experiences created by light are all for free, the happy by-products of light meeting sculpture.

Alas the complex airspaces created by the steel lines and their wall shadows *can only be seen by being there*. Sculpture's interface is reality.

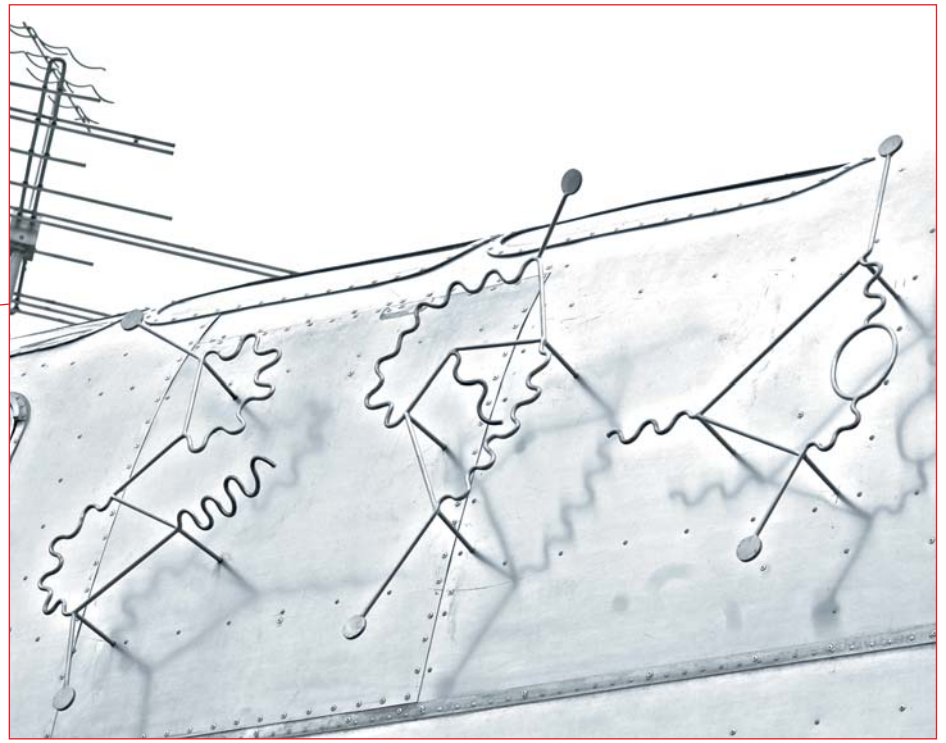
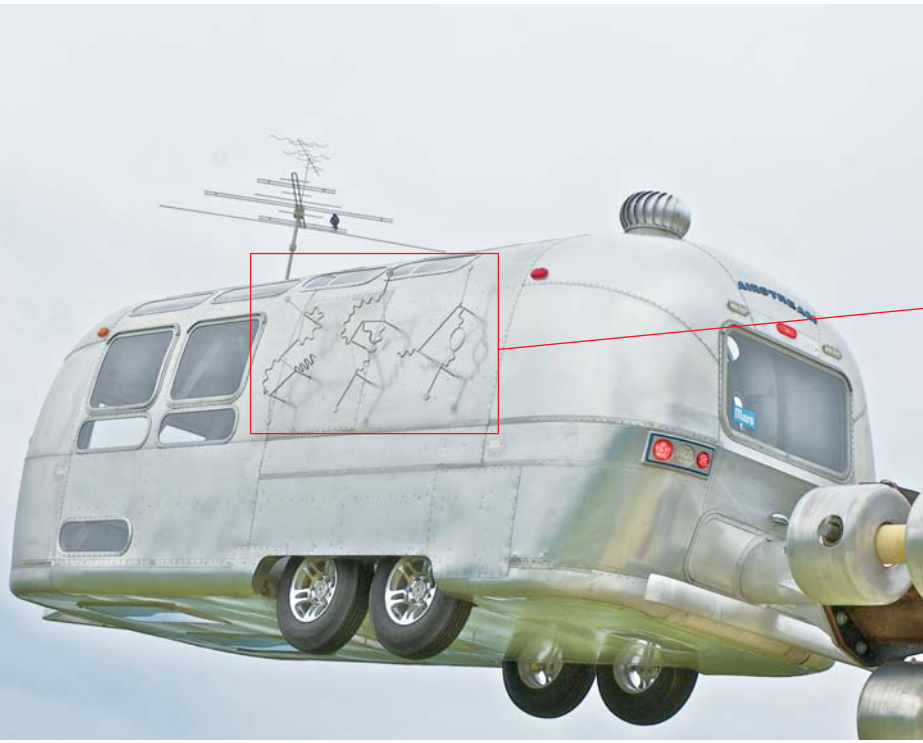
Rocket Science 2 (*Lunar Lander*) 2009
steel, aluminum, porcelain length 70 feet
or 21 meters, height 35 feet or 11 meters

Rocket Science 3 (*Airstream Interplanetary Explorer*) 2011
steel, aluminum, stainless steel, electronics length 84
feet or 26 meters, height 31 feet or 9 meters



Since Feynman diagrams describe the universal operations of Nature's laws, they can communicate throughout the universe. Both sides of the *Airstream Interplanetary Explorer* show Feynman diagrams that may well communicate with intelligent life anywhere. Better the cosmopolitan verbs of Nature's laws on spacecraft than the local proper nouns of national flags, earthly Gods and Goddesses, and government agency logos. For interplanetary exploration, better to send smart machines and emblematic Feynman diagrams than human beings and their lawn chairs, toilets, and teddie bears.

And for the cosmological entertainment of intelligent beings wherever whenever, prankish illusory violations of Nature's laws make jokes that travel well – as in the Pioneer Space Plaque redesign at right.



THE PIONEER SPACE PLAQUE: A COSMIC PRANK

Magic, the production of entertaining illusions, has an appeal quite independent of the local specifics of language, history, or culture. In vanishing objects or levitating assistants, conjurers amaze, delight, and even shock their audiences by the apparent violation of the universal laws of nature and our daily experience of those laws.

Since the principles of physics hold everywhere in the entire universe, magic is conceivably a cosmological entertainment, with the wonder induced by theatrical illusions available to and appreciated by all, regardless of planetary system. Accordingly the original plaque placed aboard the Pioneer spacecraft for extraterrestrial scrutiny billions of years from now might have escaped from its conspicuously anthropocentric gestures by showing instead the universally familiar Amazing Levitation Trick.

The Pioneer Space Plaque: A Cosmic Prank 2010 digital print, animation electronics 6.9 x 2.2 x .5 feet or 2.1 x .7 x .2 meters

Torqued Space-Time Feynman diagram

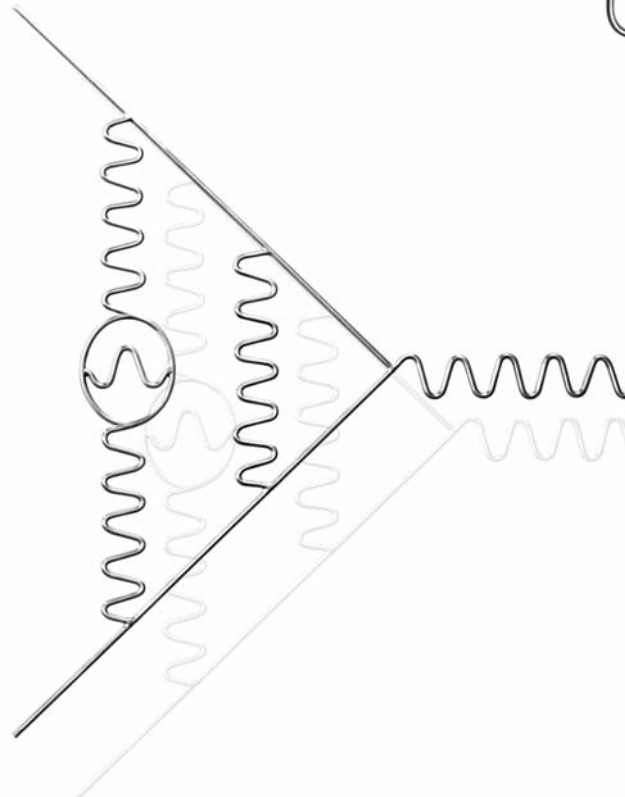
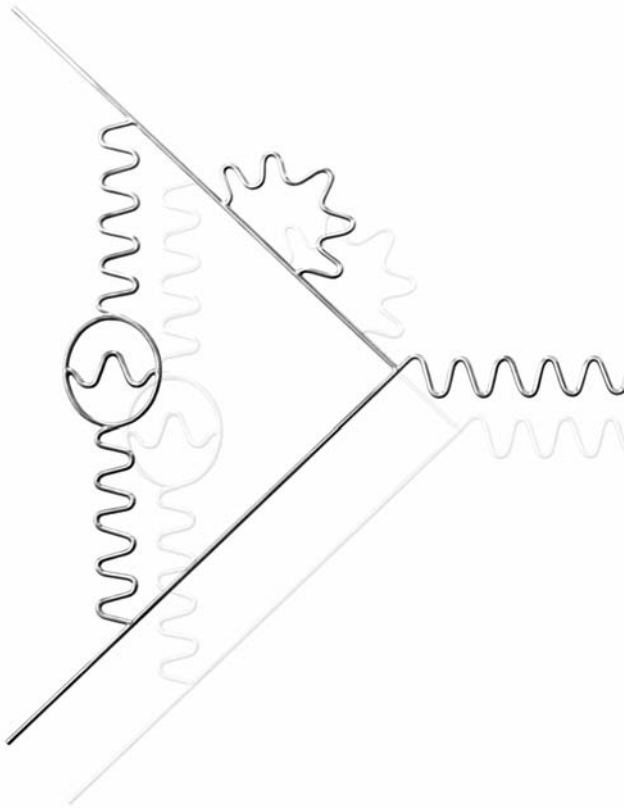
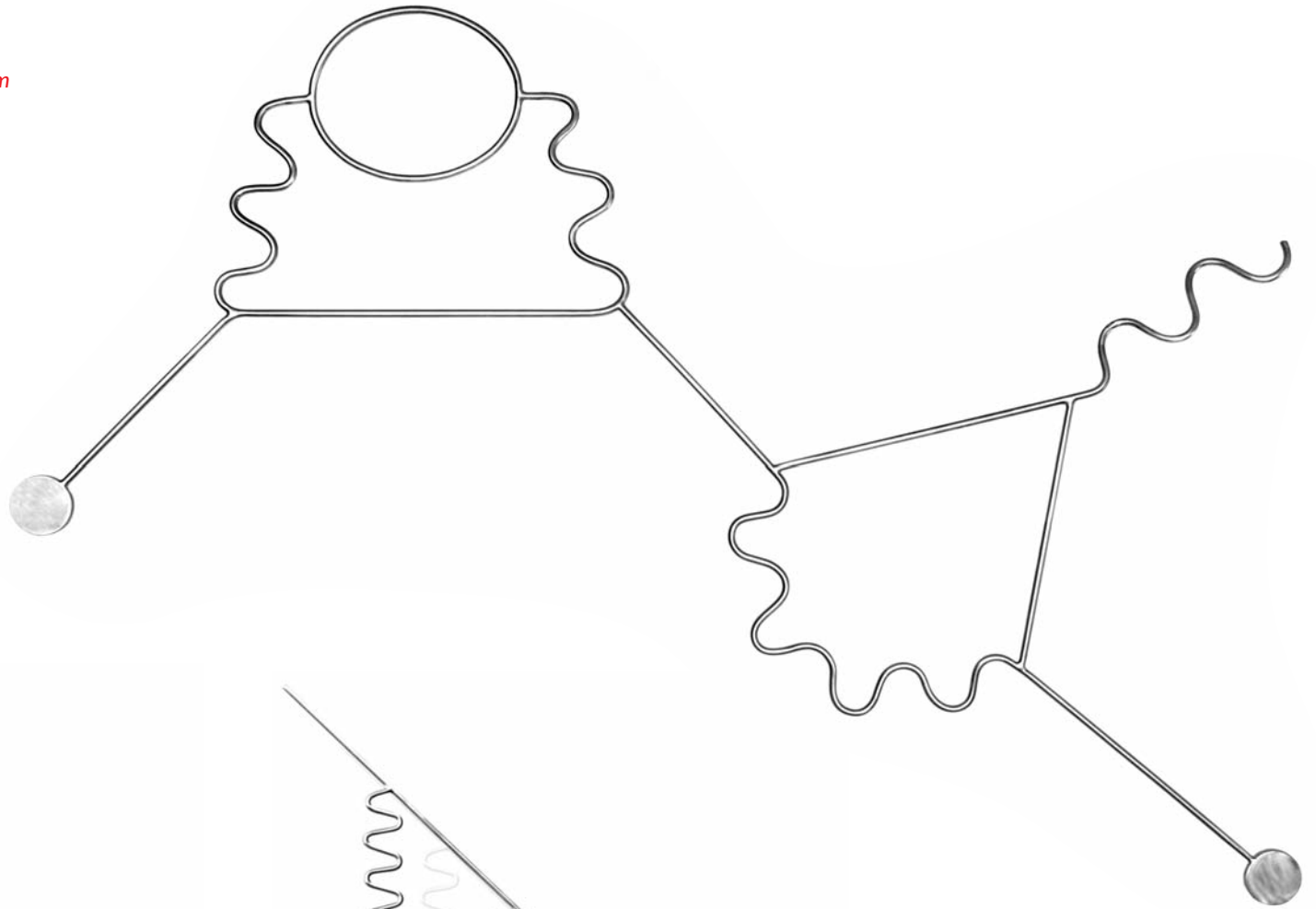
2012 stainless steel

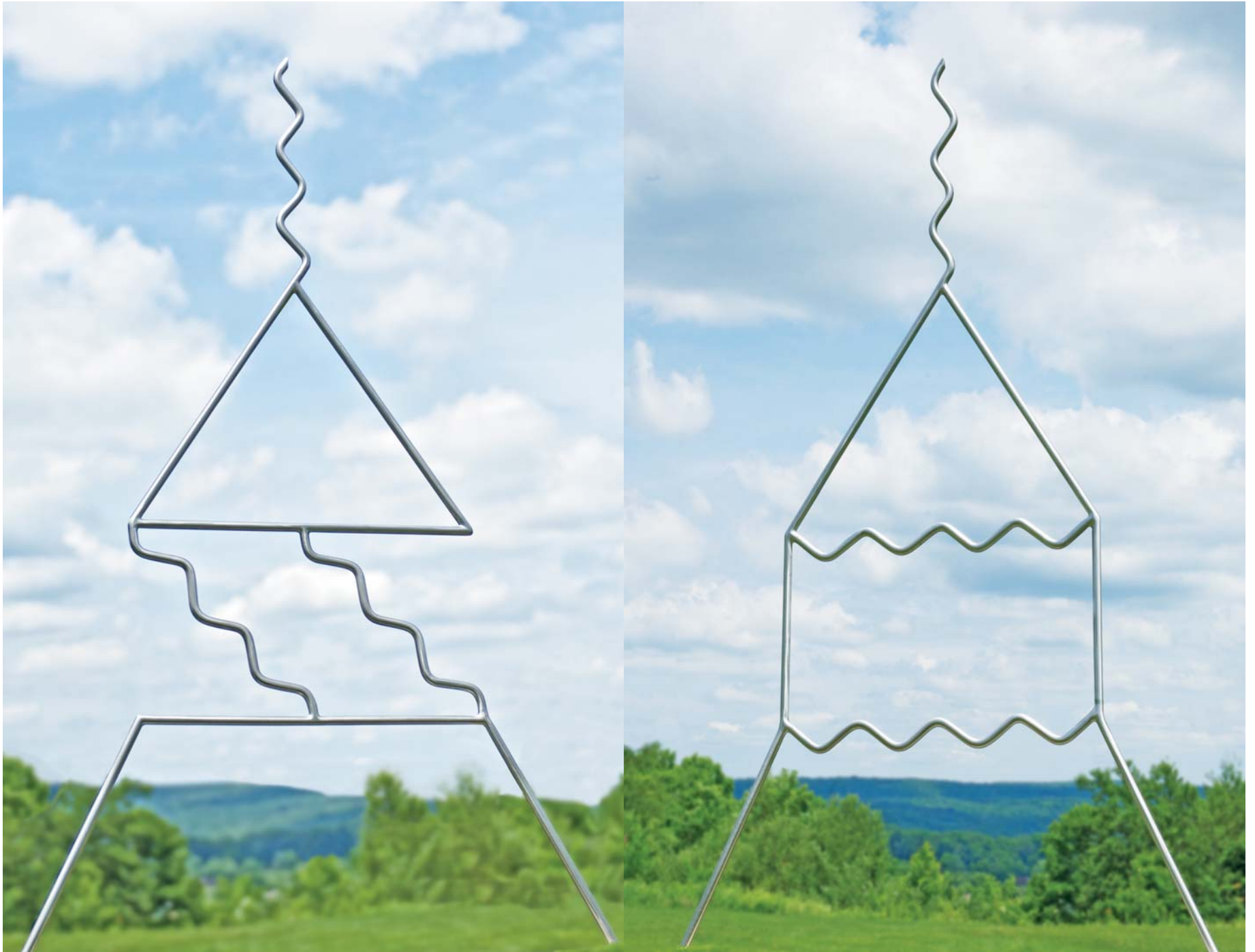
6.1 x 8.1 x 1 ft or 1.9 x 2.5 x .3 m

2 Space-Time Feynman diagrams

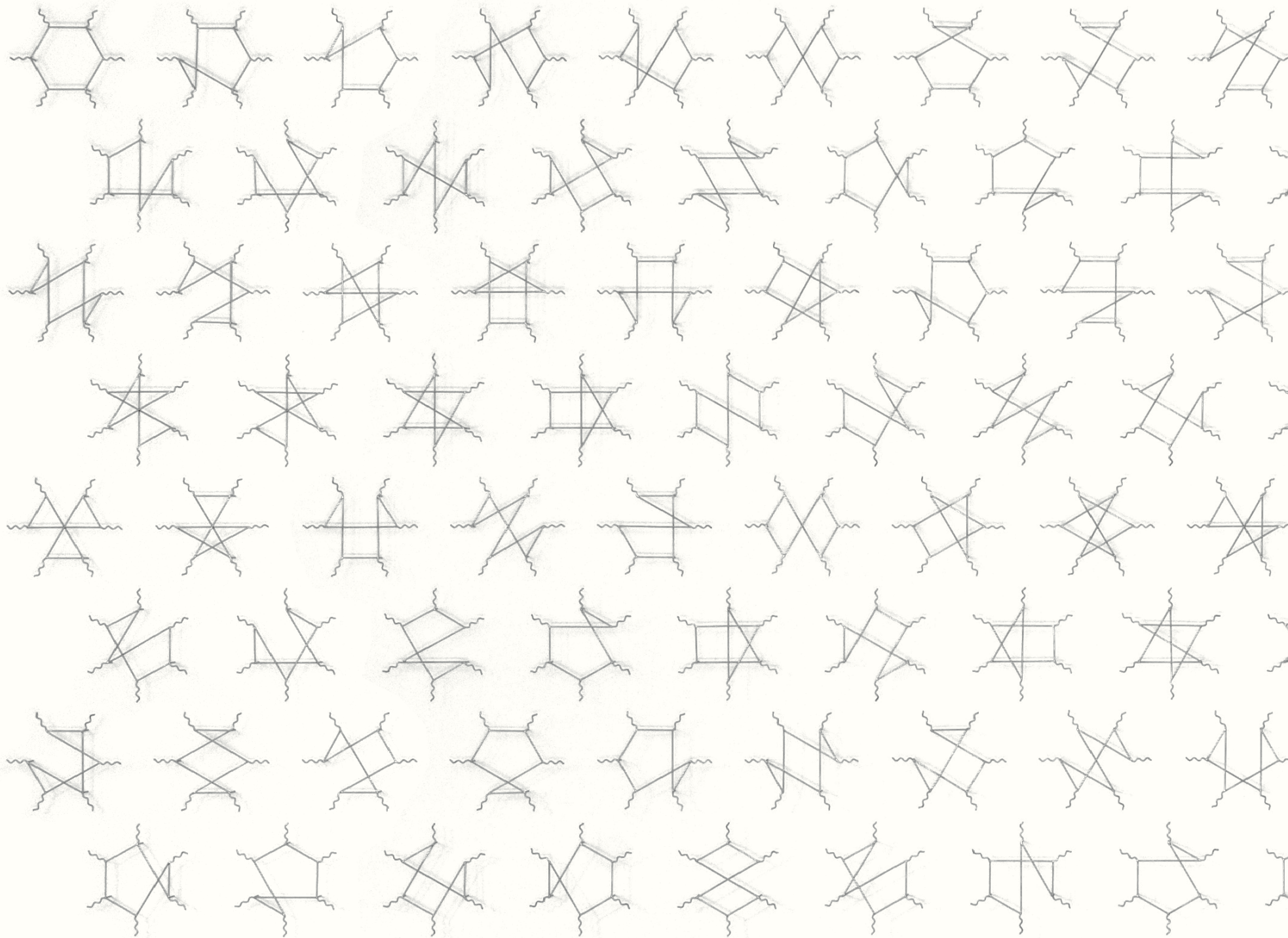
2012 stainless steel pair

8.9 x 4.5 x .3 ft or 2.7 x 1.4 x .1 m

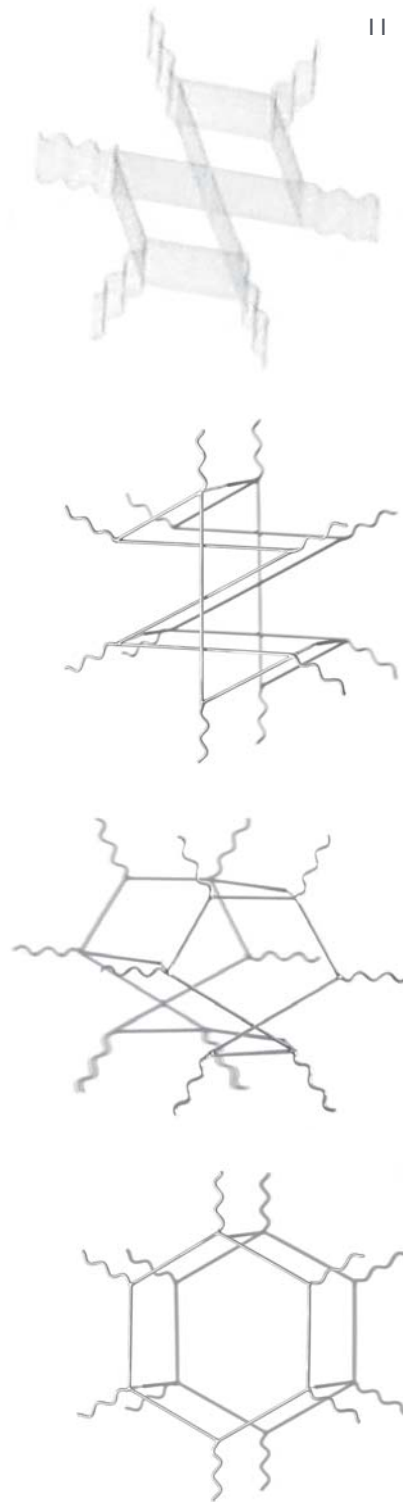
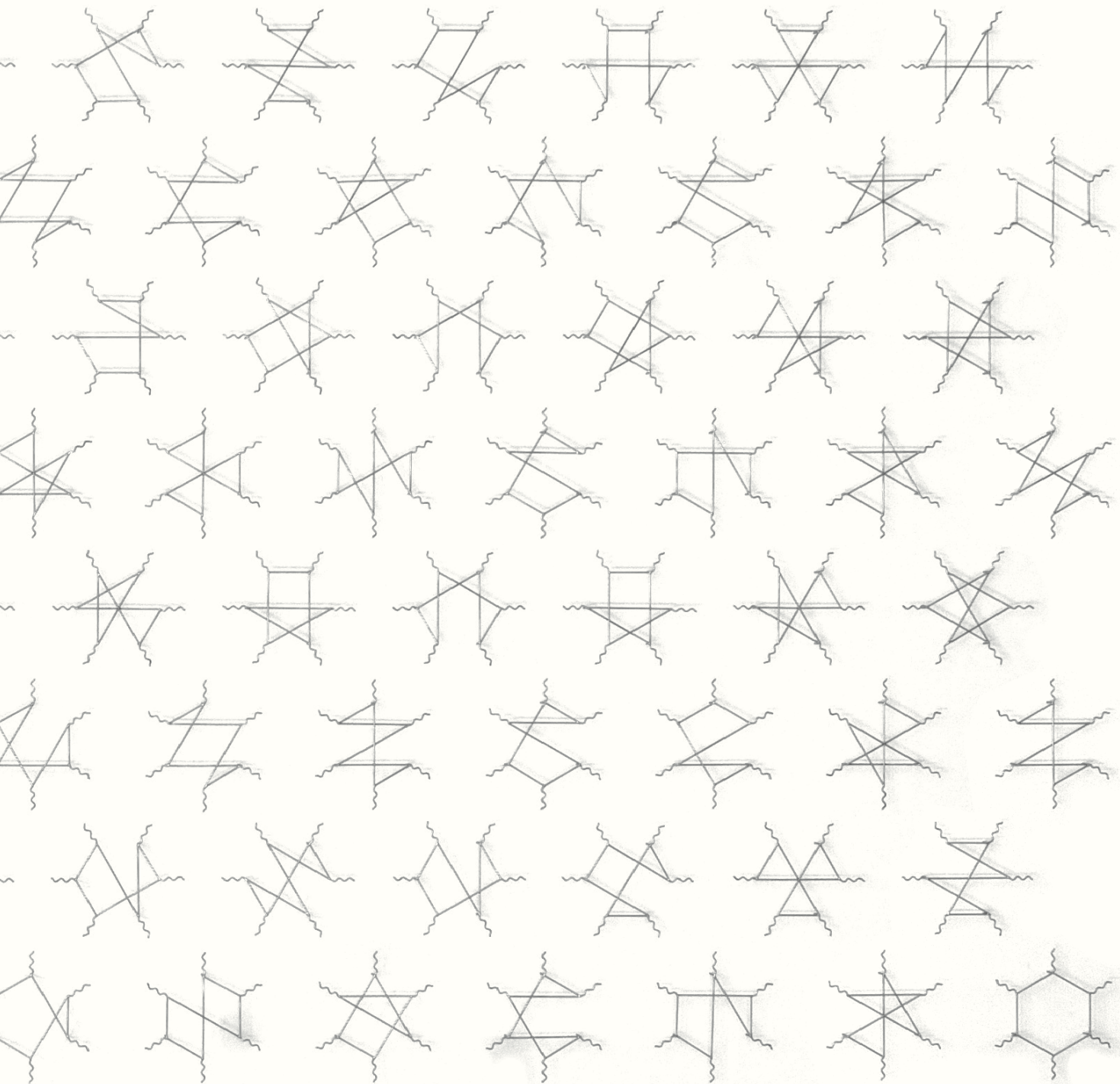




2 Space-Time Feynman Diagrams (Aztec) 2012 stainless steel 30.5 x 49 x 3.5 in or .8 x 1.2 x .1 m



All Possible 6-Photon Scattering (120 Space-Time Feynman Diagrams) 2012 stainless steel 17.5 x 7.3 x .2 feet or 5.3 x 2.3 x .1 meters



FEYNMAN DIAGRAMS

Edward Tufte, *Beautiful Evidence* (2006), pages 76-77.

WITH differentiated lines similar to maps and old electronic schematics, Richard Feynman's famous diagrams for quantum electrodynamics depict complex ideas. Based on a dictionary and elaborate syntax, the diagrams portray interactions of photons, electrons, positrons, their many colleagues and anti-colleagues by means of visual reasoning, logical enumeration, and mathematical operations. Behind the scenes, extensive calculations are at work; in 1983, the magnetic moment of an electron was computed to 12 significant digits using 900 diagrams with 100,000 terms.

Serving simultaneously as images, equations, and verbal summaries, Feynman diagrams are multimodal and thus, in practice, often modeless. For example, this double-page layout below from Martinus Veltman's *Diagrammatica: The Path to Feynman Diagrams* combines diagrams, their parallel mathematical equations, and a verbal narrative. Veltman points out that the "situation in quantum electrodynamics is more complicated," which we knew before we started. These pages below have an elegant

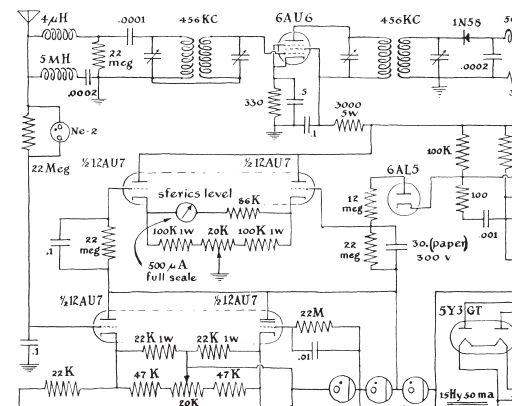


Illustration by Roger Hayward, in C. L. Stong, *The Amateur Scientist* (New York, 1960), 287.

Martinus Veltman, *Diagrammatica: The Path to Feynman Diagrams* (Cambridge, 1994), 150-151.

Now the infinity of the one loop σ self-energy diagram was absorbed in the σ mass m . That means that the σ mass is now given by $m^2 + \delta_m$. With this choice for the σ mass however there is a contribution of order g^4 hidden in the one loop diagram, see figure, where the dot signifies that for the σ mass one must take $m^2 + \delta_m$. To order g^2 one has:

$$\frac{1}{Q^2 + m^2 + \delta_m - i\epsilon} = \frac{1}{Q^2 + m^2 - i\epsilon} - \frac{1}{Q^2 + m^2 - i\epsilon} \frac{\delta_m}{Q^2 + m^2 - i\epsilon}$$

which can be pictured as shown.



Selecting then from this one loop diagram the part proportional to g^4 leads to the diagram shown, where now the cross stands for a factor $-\delta_m$. In other words, having absorbed the one loop infinity in the σ mass means that at the g^4 level we find in fact two diagrams:



The cross equals precisely minus the infinite part of the σ self-energy insertion in the first diagram. Together then they are finite; in fact, the self-energy diagram was computed before, and together the result is finite and of the form

$$F(Q) = C + \int_0^1 dx \ln(x(1-x)Q^2 + M^2)$$

where C is a constant.

divergent, and one must show that the remaining infinity is such that it can be absorbed again in the parameters of the theory, i.e., it must be a linear combination of a constant and a constant times momentum squared but nothing else.

Disentangling infinities is a rather complicated affair. One speaks of overlapping divergencies.

6.5 Quantum Electrodynamics

The situation in quantum electrodynamics is more complicated. The reason is that the electron propagator behaves as $1/k$ for large momentum rather than $1/k^2$ as the π and σ propagators in the foregoing. The divergent diagrams at the one loop level are:

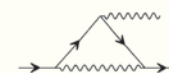
Electron self-energy, Λ



Photon self-energy, Λ^2



Electron-photon vertex, $\ln \Lambda$



Photon scattering, $\ln \Lambda$



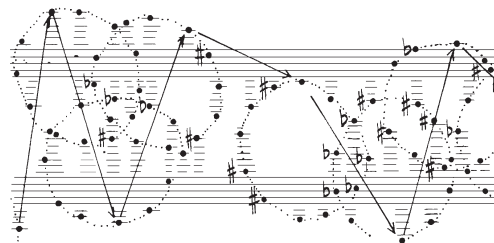
Exercise 6.5 Verify this.

The tadpole diagram is zero. The corresponding expression is:

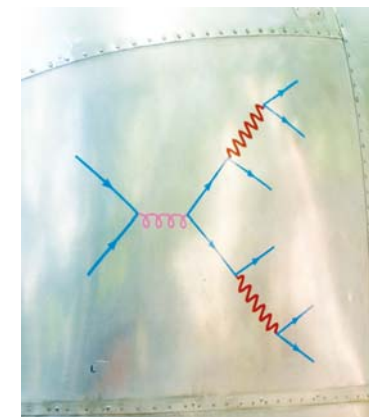
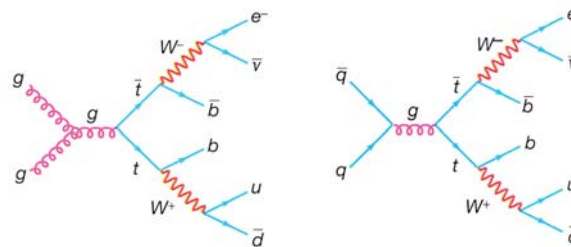
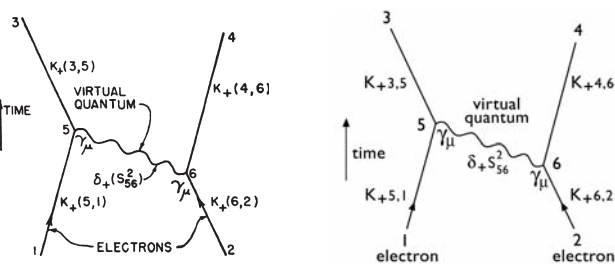
$$\int d_4p \frac{\text{Tr}\{\gamma^\mu(-i\gamma p + m)\}}{p^2 + m^2 - i\epsilon}$$



visual precision, similar to John Cage's artistic musical scores. Veltman introduces *Diagrammatica* with a note indicating that behind the graceful page layout is a thoughtful mathematical physicist and book designer: "This book is somewhat unusual in that I have tried very hard to avoid numbering the equations and figures. This [keeps] all derivations and arguments closed in themselves, and the reader needs not to have fingers at eleven places to follow an argument."¹



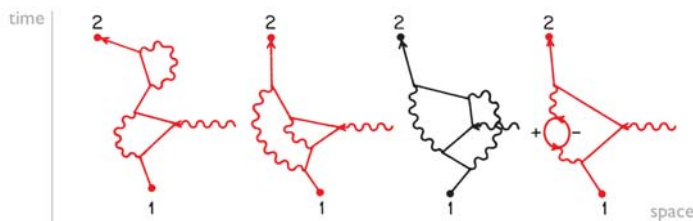
John Cage, detail from *Concert for piano and orchestra, 1958, solo for piano, 9.*



Feynman diagram on the Airstream Interplanetary Explorer.

Produced by mechanical drawing, the first published Feynman diagram (above left) commits the classic design error of *equal line weight for all visual elements*. In the original at left, naive arrows serve as pointer lines and as traces of quantum dynamics, just as dimension lines sometimes get mixed up with object lines in architectural plans. In my redesign next to the original, pointer lines prove unnecessary other than for time's arrow. More recent Feynman diagrams use color lines similar to road maps.

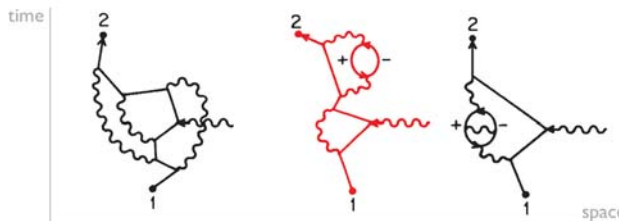
An endlessly useful strategy in analytical design is to extend the scope of a good design element: increasing the dimensionality of the space the element resides in, enhancing resolution of the element, multiplying elements, integrating the element into various displays. Such is the history of Feynman diagrams, as below on 2 space-time grids multiple quantum dancers move about, described by Feynman's words beneath:



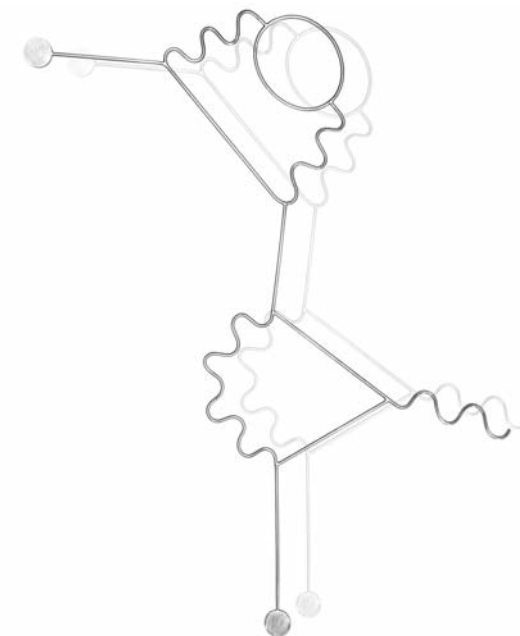
Laboratory experiments became so accurate that further alternatives, involving four extra couplings (over all possible intermediate points in space-time), had to be calculated, some of which are shown here. The alternative on the right involves a photon disintegrating into a positron-electron pair, which annihilates to form a new photon, which is ultimately absorbed by the electron.

¹ Veltman, *Diagrammatica*, xii.

Far left: R. P. Feynman, "Space-time approach to quantum electrodynamics," *Physical Review*, 76 (1949), 776. Color diagrams: DØ Collaboration, "A precision measurement of the mass of the top quark," *Nature*, 429 (10 June 2004), 638-641. Below: Richard P. Feynman, *QED: The Strange Theory of Light and Matter* (Princeton, 1985), 117-118. For history, see David Kaiser, *Drawing Theories Apart: The Dispersion of Feynman Diagrams in Postwar Physics* (Chicago, 2005).



Calculations are presently going on to make the theoretical value even more accurate. The next contribution to amplitude, which represents all possibilities with six extra couplings, involves 70 diagrams, 3 of which are shown here. As of 1983, the theoretical number was 1.00115965246, and the experimental number was 1.00115965221.



Shown in red at left, 4 Feynman diagrams were translated into stainless steel sculptures.

The Feynman-Tufte Principle

A visual display of data should be simple enough to fit on the side of a van By MICHAEL SHERMER

I had long wanted to meet Edward R. Tufte—the man the *New York Times* called “the da Vinci of data” because of his concisely written and artfully illustrated books on the visual display of data—and invite him to speak at the Skeptics Society science lecture series that I host at the California Institute of Technology. Tufte is one of the world’s leading experts on a core tool of skepticism: how to see through information obfuscation.

But how could we afford someone of his stature? “My honorarium,” he told me, “is to see Feynman’s van.”

Richard Feynman, the late Caltech physicist, is famous for working on the atomic bomb, winning a Nobel Prize in Physics, cracking safes, playing drums and driving a 1975 Dodge Maxivan adorned with squiggly lines on the side panels. Most people who saw it gazed in puzzlement, but once in a while someone would ask the driver why he had Feynman diagrams all over his van, only to be told, “Because I’m Richard Feynman!”



EDWARD R. TUFTE, master of design analysis, poses next to a Feynman diagram on Feynman’s van depicting the interaction of photons and electrons.

Feynman diagrams are simplified visual representations of the very complex world of quantum electrodynamics (QED), in which particles of light called photons are depicted by wavy lines, negatively charged electrons are depicted by straight or curved nonwavy lines, and line junctions show electrons emitting or absorbing a photon. In the diagram on the back door of the van, seen in the photograph above with Tufte, time flows from bottom to top. The pair of electrons (the straight lines) are moving toward each other. When the left-hand electron emits a photon (wavy-line junction), that negatively charged particle is deflected outward left; the right-hand electron reabsorbs the photon, causing it to deflect outward right.

Feynman diagrams are the embodiment of what Tufte teaches about analytical design: “Good displays of data help to reveal knowledge relevant to understanding mechanism,

process and dynamics, cause and effect.” We see the unthinkable and think the unseeable. “Visual representations of evidence should be governed by principles of reasoning about quantitative evidence. Clear and precise seeing becomes as one with clear and precise thinking.”

The master of clear and precise thinking meets the master of clear and precise seeing in what I call the Feynman-Tufte Principle: a visual display of data should be simple enough to fit on the side of a van.

As Tufte poignantly demonstrated in his analysis of the space shuttle *Challenger* disaster, despite the 13 charts prepared for NASA by Thiokol (the makers of the solid-rocket booster that blew up), they failed to communicate the link between cool temperature and O-ring damage on earlier flights. The loss of the *Columbia*, Tufte believes, was directly related to “a PowerPoint festival of bureaucratic hyperrationalism” in which a single slide contained six different levels of hierarchy (chapters and subheads), thereby obfuscating the conclusion that damage to the left wing might have been significant. In his 1970 classic work *The Feynman Lectures on Physics*, Feynman covered all of physics—from celestial mechanics to quantum electrodynamics—with only two levels of hierarchy.

Tufte codified the design process into six principles: “(1) documenting the sources and characteristics of the data, (2) insistently enforcing appropriate comparisons, (3) demonstrating mechanisms of cause and effect, (4) expressing those mechanisms quantitatively, (5) recognizing the inherently multivariate nature of analytic problems, (6) inspecting and evaluating alternative explanations.” In brief, “information displays should be documentary, comparative, causal and explanatory, quantified, multivariate, exploratory, skeptical.”

Skeptical. How fitting for this column, opus 50 for me, because when I asked Tufte to summarize the goal of his work, he said, “Simple design, intense content.” Because we all need a mark at which to aim (one meaning of “skeptic”), “simple design, intense content” is a sound objective for this series. ■

Michael Shermer is publisher of *Skeptic* (www.skeptic.com) and author of *Science Friction*.



In addition to the *All Possible Photons* wall sculptures, Edward Tufte has constructed 70 large-scale landscape artworks, including *Twigs* (series of 6), *Escaping Flatland* (10), *Rocket Science* (3), and *Megaliths of Silence* (40 large stone works so far).

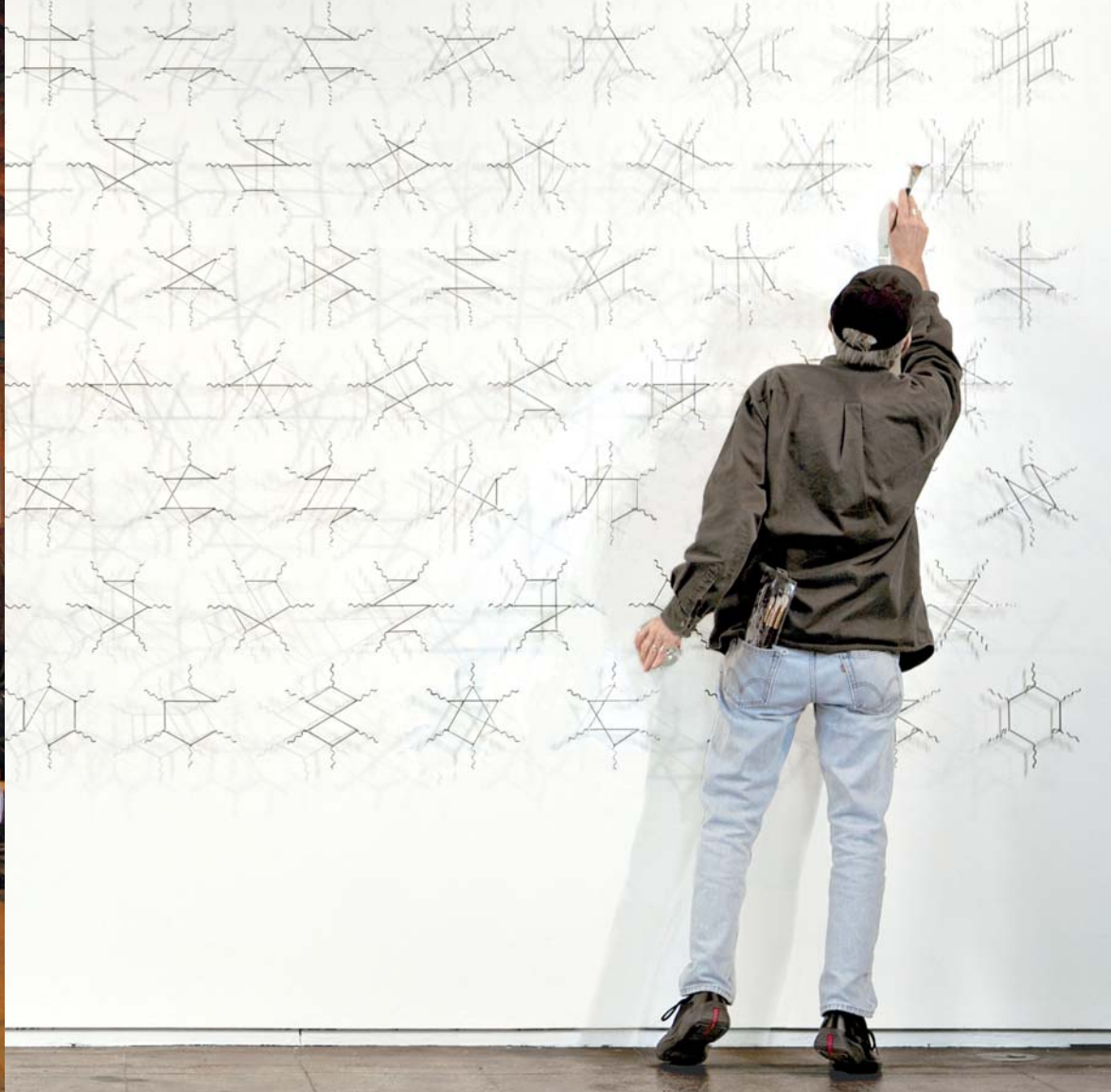
ET wrote, designed, and self-published 4 books on data science, visual thinking, and analytical design: *The Visual Display of Quantitative Information*, *Envisioning Information*, *Visual Explanations*, and *Beautiful Evidence*. *The New York Times* described ET as the “da Vinci of data” and *Business Week* as the “Galileo of graphics.” He served as a professor at Princeton University and Yale University for 33 years, and in 2010 was appointed by the President to the Recovery Independent Advisory Panel.

He has received 7 honorary degrees, and is a fellow of the American Academy of Arts and Sciences, the Guggenheim Foundation, and the Center for Advanced Study in the Behavioral Sciences.

ET's sculpture fields are on 266 acres in Connecticut. Accounts of the work are at www.tufte.com.

Feynman diagrams painted by Richard Feynman on his van in 1975. Garage storage, Long Beach, California, 2004.





EDWARD TUFTE ALL POSSIBLE PHOTONS

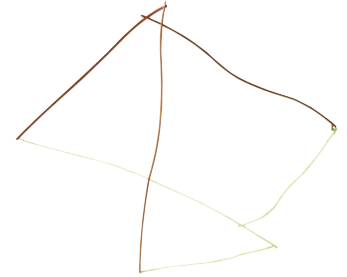
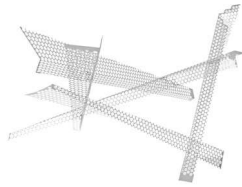
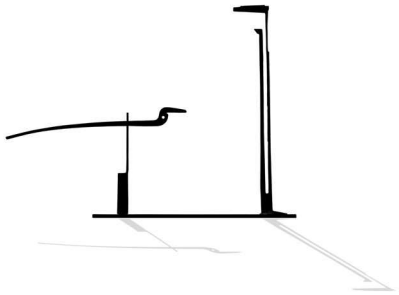
THE CONCEPTUAL AND COGNITIVE ART OF FEYNMAN DIAGRAMS

 MODERN

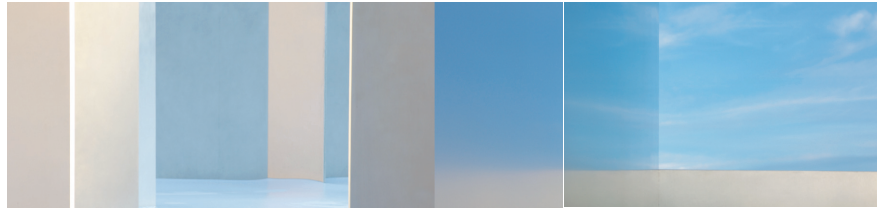
ET MODERN 547 WEST 20TH STREET NEW YORK 10011

CORNER OF WEST 20TH STREET + 11TH AVENUE WWW.TUFTE.COM 212.206.0300

SEEING AROUND EDWARD TUFTE



FORM AND SCALE



COLOR IN SPACE AND TIME



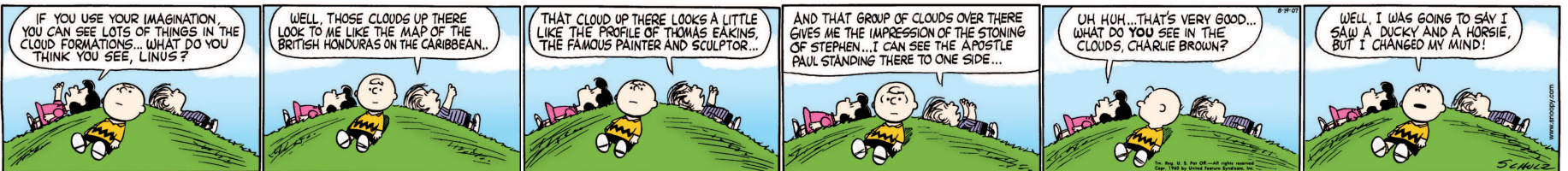
ABSTRACT SCULPTURE MEETS THE LAND

ANIMALS, SHADOWS, DAPPLES, REFLECTIONS MOVING IN SPACE AND TIME

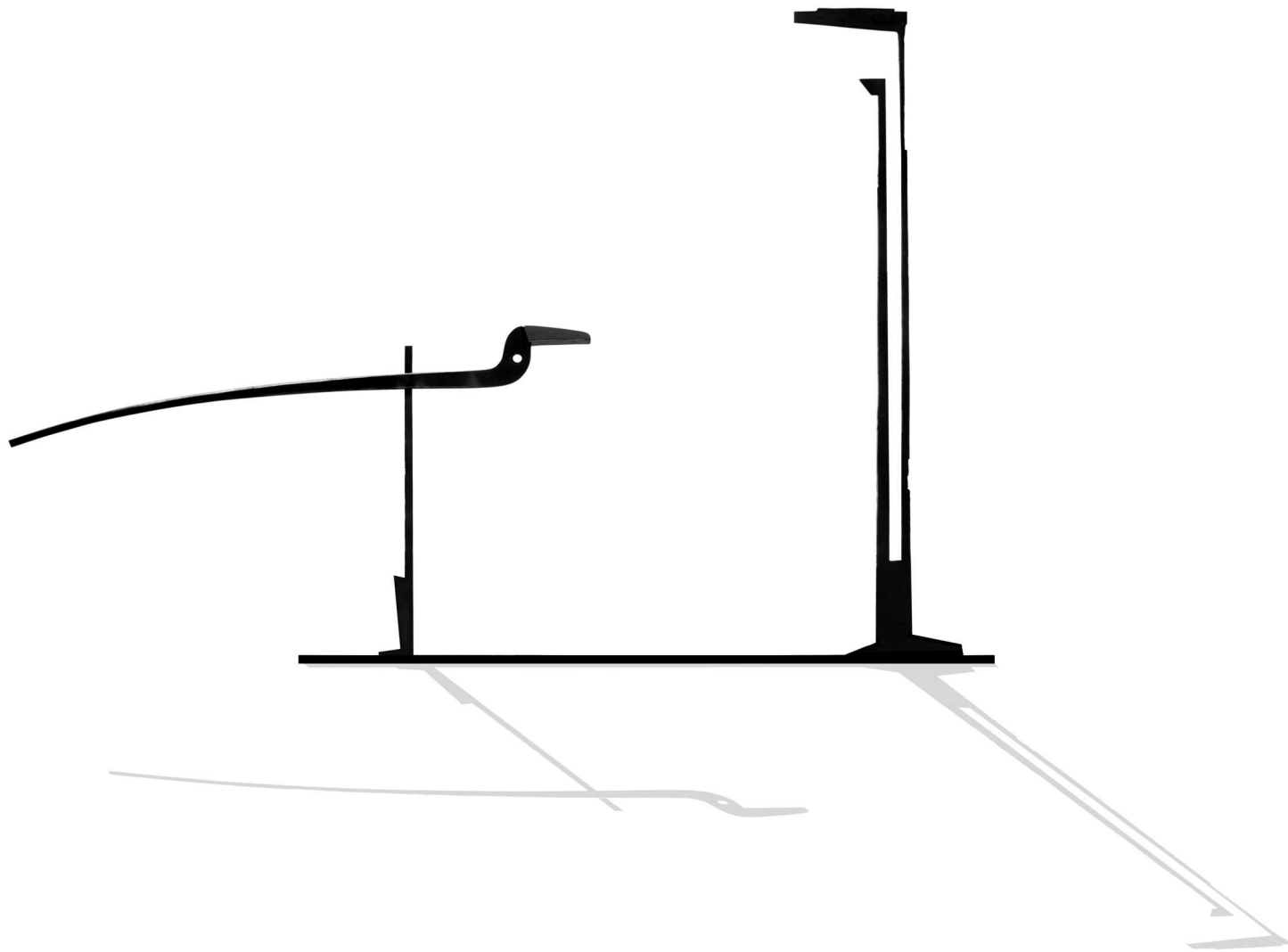


AIRSPACE, LANDSCAPE

PRODUCTION AND INSTALLATION: STEEL, GRAVITY, HOPE



SEE NOW . . . WORDS LATER



FORM AND SCALE

Tong Bird of Paradise 2008
steel height 20 feet or 6 meters



Larkin's Twig 2004

steel height 32 feet or 10 meters

footprint 59 x 66 x 70 feet or 18 x 20 x 21 meters

Sculptures are artworks that cast shadows. The fundamental being of sculptural form is its residence in Nature's three-dimensional reality and the visual multiplicities that result from such residence.

Sculptural form is what the work is about, the 3D relationships within the piece and with its airspace. Good form is somehow the coherent and varying interplay among these relationships. Good form resides fully in all three spatial dimensions, has a wholeness and coherence, is often vigorously asymmetric, varies gracefully from varying points of view, repays study, and is intriguing, bright-eyed, fresh.

The *physical size* of a sculpture is what it is. But its *scale* is relational, tied to everything in the neighborhood: the artwork's size, form, and airspace; the context (city plaza, valley, meadow, museum room, intimate garden); and the viewer's position relative to the piece. *Rocket Science*, physically large, appears immense and disorienting to nearby observers on the land, delightful to children, and not all that big a presence compared to nearby land, trees, horizon hills.

Thoughtful scaling is inherent to creating large artworks. Sculptors decide whether to begin with a small model or, better, to work the piece at full size. Small models scale up in non-obvious ways, relative to themselves and to their surroundings. What works small doesn't necessarily work medium or large. Some big outdoor pieces lack scale coherence and look like tiny models immensely enlarged. Frank Gehry works with small architectural models *at several sizes* to avoid designing a perfect toy building. Mies van der Rohe suggested viewing objects at sizes much smaller *and* much larger than actual size.

Rocket Science was designed at full scale. Big cranes temporarily held together rough drafts of the artwork. A small cardboard model later helped engineer the legs and locate the piece. Photographs and videos were shot at several scales. Full-size mock-ups, small models, computer animations, sketches, confections, photographs, videos: *whatever it takes* to see, understand, and construct artwork.

In scale it's incredibly important outdoors to have a presence underneath the sky, underneath the sun, and I never, ever think of competing with what nature does in any way. But I do try to hold my own with the surroundings, to have a presence in the context of the surroundings.

Ursula von Rydingsvard



Airspace 2009 aluminum height 15 feet or 4.6 meters



Rocket Science I (Spacecraft) 2007
steel height 32 feet or 10 meters
48,000 lbs or 21,800 kgs



Birds 11, 14, 15, 16, 18 2005-2006

anodized aluminum variable sizes

~60 x 40 x 80 inches or ~1.5 x 1.0 x 2.0 meters



Towers 2006 (detail) stainless steel
height 7.5 feet or 2.3 meters

COLOR IN SPACE AND TIME

Outdoor stainless steel artworks paint beautiful and subtle color fields by borrowing, altering, absorbing, and reflecting nature's light. Such artworks are both a cause and effect of light. Color fields vary with ambient light, the stainless steel surfacing, and the viewer's position relative to the piece. The happy consequence is a multiplicity of dynamic color-field paintings in three-space.

David Smith's description of light on stainless steel is a reminder of Impressionist painting and, in particular, Claude Monet's *Haystack* paintings, which create intense color experiences by portraying and amplifying the changing colors of the seasons, hours, weather, open air surroundings.


Sequential photographs document and reveal color changes on stainless steel. But photographs are flat and still representations, with a limited dynamic range compared to the eye. To see the luscious 3D color, only being there with the artwork and its atmospheric will do. Sometimes even the airspace around the piece glows with color.

I like outdoor sculpture and the most practical thing for outdoor sculpture is stainless steel, and I make them and I polish them in such a way that on a dull day, they take on the dull blue, or the color of the sky in the late afternoon sun, the glow, golden like the rays, the colors of nature. And in a particular sense, I have used atmosphere in a reflective way on the surfaces. They are colored by the sky and the surroundings, the green or blue of water. Some are down by the water and some are by the mountains. They reflect the colors. They are designed for outdoors.

David Smith

Stainless Steel Engraving 7 2006 rotation under ambient outdoor light



The image consists of two vertical panels showing a sculpture against a clear blue sky. The sculpture is a thin, curved stainless steel rod with three teardrop-shaped elements attached to it. The left panel shows the sculpture from a low angle, looking up, with the rod extending from the bottom left towards the top right. The right panel shows the sculpture from a similar low angle, but the rod is more vertical, extending from the bottom right towards the top left. The teardrop shapes are positioned at regular intervals along the curve of the rod.

Zerlina's Smile 2008
stainless steel
height 18 feet or 5.5 meters



Escaping Flatland 10 2002 stainless steel



Bouquet 4 2006 stainless steel

Small changes in surface texture, where the light hits the metal, yield remarkable changes in reflected light. Sculptors grind, scrape, brush, rust, polish, scratch, and corrode metal to rearrange its surface atoms so as to respond elegantly to changes in outdoor light.

At far left, the surface for *Escaping Flatland* was made by *double-action grinders*, which produce a soft neutral variation when a viewer directly faces a stainless steel plate. But that same surface, seen from a flat raking angle, is a reflective mirror! Light from stainless steel depends on the character of ambient light, the micro-textures of reflecting surfaces, and the observer's relation to both. Active viewers see more. And active light, such as the sun, makes more to see.

Single-action grinding produces metal brushstrokes that read as deeply three-dimensional, which vary with the incident light. Like dappled light, these *anisotropic reflections* (an-i-so-tro-pic) appear everywhere once you know about them. They occur naturally in light from water waves, hair highlights, metal surfaces—and artificially in computer simulations of real things.

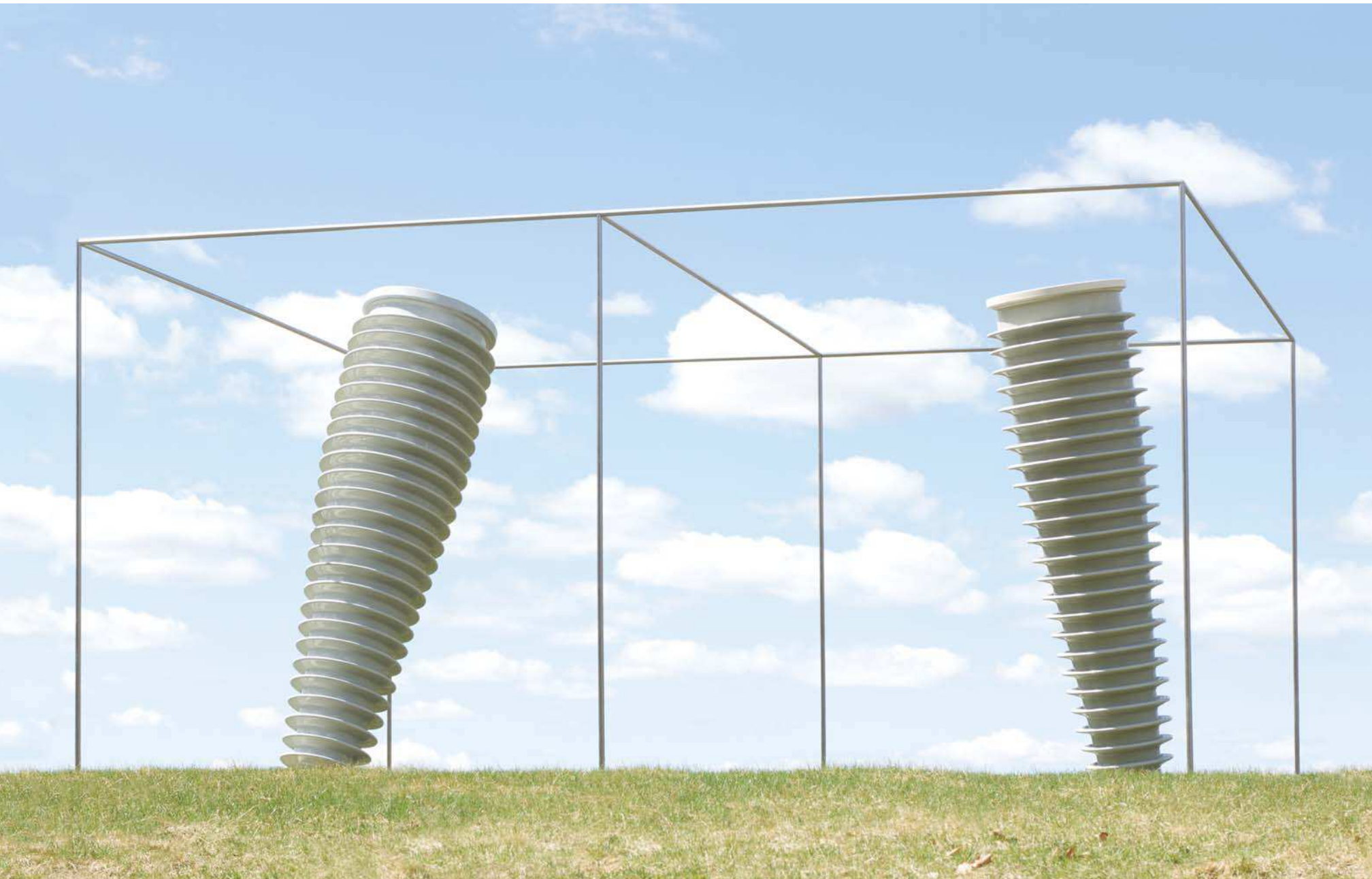
Above, single-action grinding creates vivid gestural brushstrokes in the stainless steel that articulate and reflect nearby light. This surface plays with color's hue, saturation, and value—but mirrors very little shape information from nearby objects. Small changes in viewing angle then paint endless color fields, thanks to anisotropic reflections.



Escaping Flatland 6, 7, 10 2001-2002 stainless steel height 12 feet or 3.7 meters



AIRSPACE, LANDSCAPE



Dear Leader 2006 stainless steel, porcelain 21.5 x 14 x height 10.5 feet or 6.6 x 4.3 x height 3.2 meters

On the flatlands of paper and computer screens, the interplay of figure and background is universally fundamental to visual thinking. Similarly, sculptures generate *airspace*s, 3D volumes of space near the work. During design and analysis, sculptors assess the airspaces surrounding the piece, as well as the multiple silhouettes generated by the piece against a background of sky and land.

Outdoor installation artists are sometimes able to shape the land around the artwork and thereby control airspaces and silhouettes produced by the piece on the land. Installation artworks are thus fortunate. Most other sculptures must perform at pre-assigned sites—museum rooms, corporate or public plazas, client's backyards. Now and again, arbitrary site conditions provoke the production of theatrical, lonely, competitive pieces that seek to defend themselves against nearby visual atrocities.

As viewers walk around the landscape and change their relation to the piece, they see various airspaces and silhouettes at various distances. 3D artworks have many points of view, 2D flats few.



In every clear concept of the nature of vision and in every healthy approach to the spatial world, this dynamic unity of figure and background has been clearly understood. Lao Tzu showed such grasp when he said: "A vessel is useful only through its emptiness. It is the space opened in a wall that serves as a window. Thus it is the nonexistent in things which makes them serviceable." Eastern visual culture has a deep understanding of the role of empty space in the image. Chinese and Japanese painters have the admirable courage to leave empty large paths of their picture-surface so that the surface is divided into unequal intervals which, through their spacing, force the eye of the spectator to movements of varying velocity in following up relationships, and thus create the unity by the greatest possible variation of surface. Chinese and Japanese calligraphy also have a sound respect for the white interval. Characters are written in imaginary squares, the blank areas of which are given as much consideration as the graphic units, the strokes. Written or printed communications are living or dead depending upon the organization of their blank spaces. A single character gains clarity and meaning by an orderly relationship of the space background which surrounds it. The greater the variety and distinction among respective background units, the clearer becomes the comprehension of a character as an individual expression or sign.

Gyorgy Kepes

Sculpture is an art of the open air. Daylight, sunlight, is necessary to it and, for me, its best setting and complement is nature. I would rather have a piece of my sculpture put in a landscape, almost any landscape, than in, or on, the most beautiful building I know.

Henry Moore



Spring Arcs 2004 stainless steel footprint 12 x 67 feet or 3.7 x 20.4 meters



Rocket Science 2 (Lunar Lander) 2009 steel, aluminum, porcelain length 70 feet or 21.3 meters, height 35 feet or 10.7 meters



SHADOWS AND DAPPLES MOVING IN SPACE AND TIME

As the Earth rotates, sculptures cast moving shadows on themselves and the landscape. Shadows change with every change in light to form new art. Time-lapse videos compress hours to preserve minutes of cinematic shadows sometimes as rich and complex as the artwork itself. New shadows make new scenes, a multiplicity unavailable in flat art.

Dappled light appears on artworks when sunlight is filtered through tree leaves. Dapples occur not because tree leaves have elliptical holes in them but rather because the leaves combine to make many tiny pinhole cameras, which then project many images of the Sun's surface directly onto Earth's sculptures. Knowing this, you'll never see dapples the same way again. On the piece at far right, every dapple is a direct image of the sun. And if the sun has an unusually large sunspot, it would be seen (upside down) within every dapple, as Galileo discovered in 1613.

When a breeze animates tree leaves, dapples flow back and forth over the artwork's surface. Like dappled light on a garden path, gently shifting dappled light on stainless steel is beautiful. Alert viewers will *look back towards the generating sources* of dapples and shadows – here, the nearby trees where each bright spot is the generating source of a dapple:

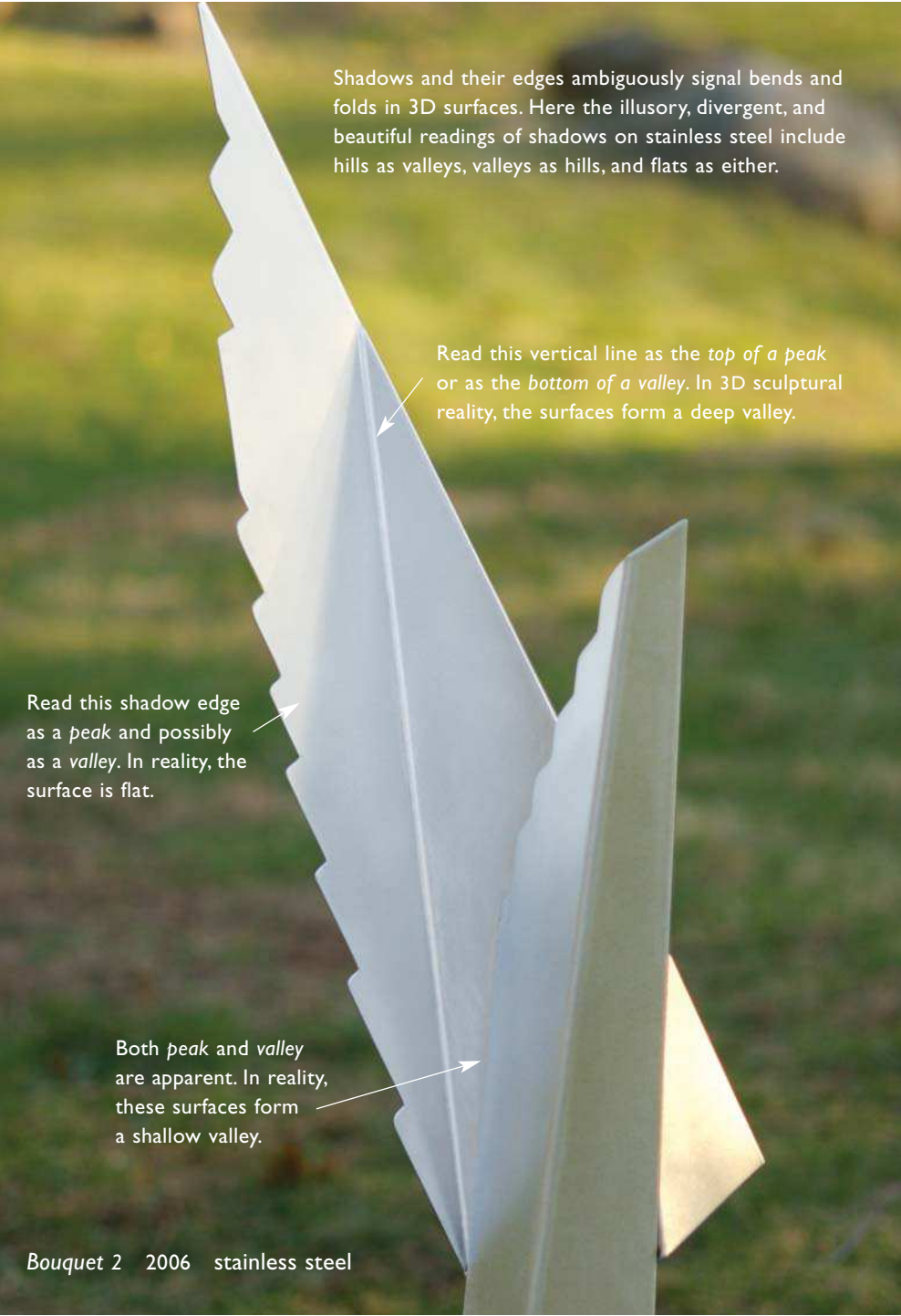




Airspace Dancers 2006 stainless steel

Escaping Flatland 7 2002 stainless steel





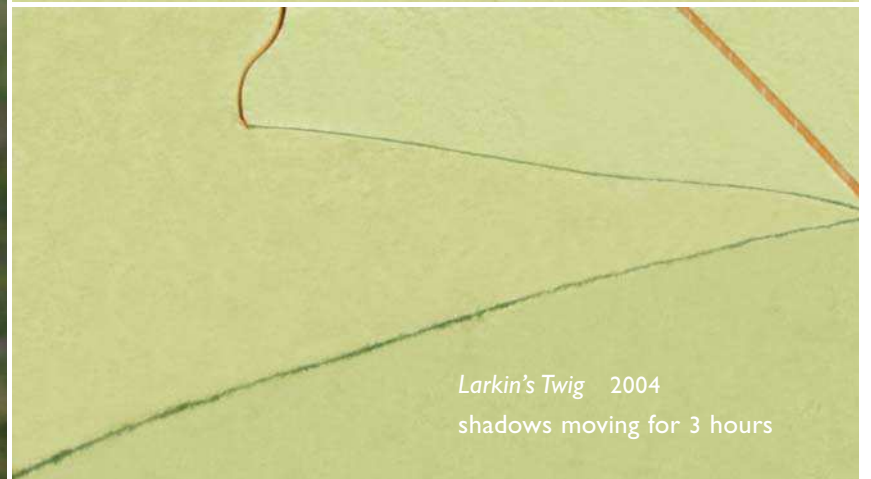
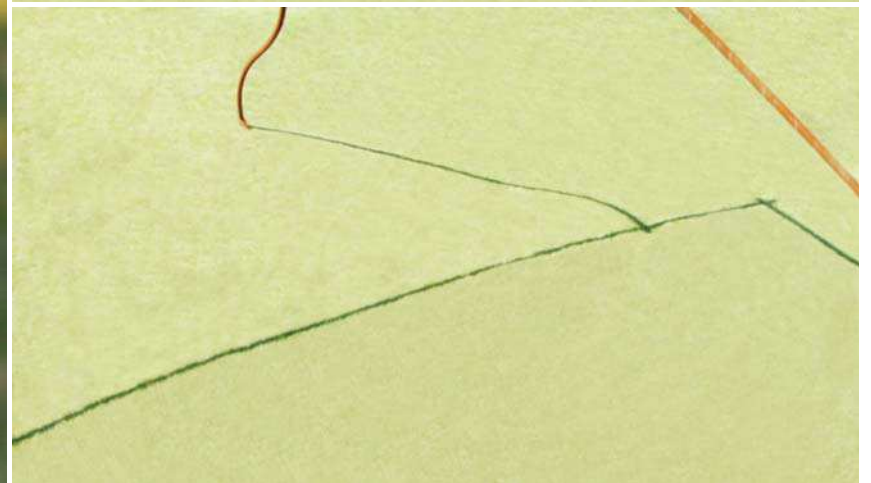
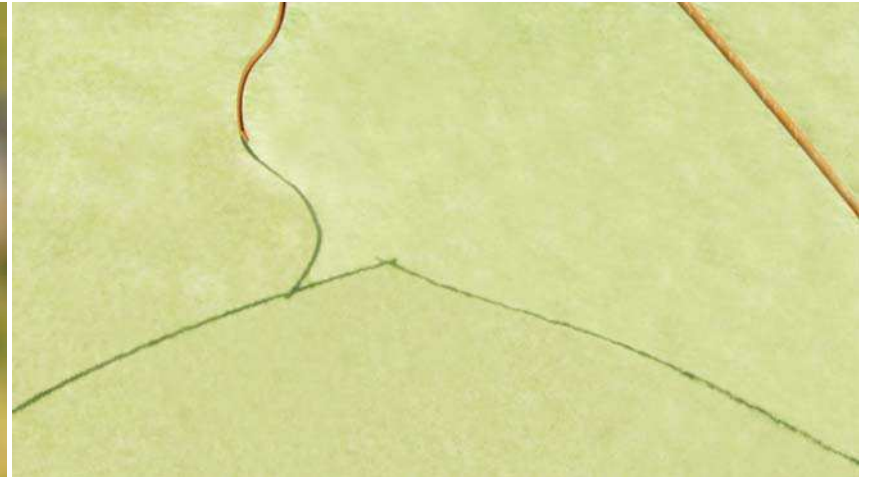
Shadows and their edges ambiguously signal bends and folds in 3D surfaces. Here the illusory, divergent, and beautiful readings of shadows on stainless steel include hills as valleys, valleys as hills, and flats as either.

Read this vertical line as the *top of a peak* or as the *bottom of a valley*. In 3D sculptural reality, the surfaces form a deep valley.

Read this shadow edge as a *peak* and possibly as a *valley*. In reality, the surface is flat.

Both *peak* and *valley* are apparent. In reality, these surfaces form a shallow valley.

Bouquet 2 2006 stainless steel



Larkin's Twig 2004
shadows moving for 3 hours



Rocket Science 2007
shadows of spacecraft visitors

ABSTRACT SCULPTURE MEETS THE LAND

Sculptural mass in three-space is an inevitable prisoner of gravity; the Earth resists, supports, aligns the piece. Representing the physical and symbolic transition from earth-flatland to sculpture-spaceland, the intersection of land and sculpture announces the beginning of art. Most big abstract sculptures sit directly on grassy land rather than on traditional pedestals.

At higher levels of modern achievement, sculpture-on-a-pedestal has faded away. Pedestalized Great Leaders appear as corny hackwork—and, for 50 years now, theatrical depedestalizations of Great Leader statues have become the performance art of radical political change.

The intersection of sculpture and land activates the nearby airspace, begins the flow of shadows from the piece, and serves as a pivot for shadows flowing around the piece as the Earth rotates. This lively intersection should be uncluttered and pristine (and kept that way during land maintenance)—so that the artwork, its edgeline resting on the land, the airspace, and shadows all play together elegantly.

Millstone 5, poised on the land



Dear Leader, shadows on ground, perspective box reflecting sunlight

The biggest break in the history of sculpture in the twentieth century was to remove the pedestal. The historical concept of placing sculpture on a pedestal was to establish a separation from the behavioral space of the viewer. "Pedestalized" sculpture invariably transfers the effect of power by subjugating the viewer to the idealized, memorialized or eulogized theme. As soon as art is forced or persuaded to serve alien values it ceases to serve its own needs. To deprive art of its uselessness is to make other than art.

Richard Serra

In sculpture fields, the landscape's common surface ties pieces together. As viewers walk around—*ambulatory seeing* in contrast to *fixed-position seeing*—everything is moving one way or another. What then should be the *pace* of that movement? With big landscape pieces, changes are often seen better when the pace is *faster* than walking, perhaps 2 revolutions per minute while circling around the piece at an appropriate distance, followed by a quiet stroll up-close. This is similar to varying the frame rate in making movies, from time-lapse to slow motion.

Escaping Flatland, ice reflections

Larkin's Twig, ice shadows





Millstone 3 and Millstone 1 2003
mild steel height 13.7 feet or 4.2 meters
weight 13,000 lbs or 5,900 kg each



Larkin's Twig 2004

steel height 32 feet or 10 meters

footprint 59 x 66 x 70 feet or 18 x 20 x 21 meters

Escaping Flatland 5-10 2001-2002

stainless steel width 50 feet or 15 meters

The Aldrich Contemporary Art Museum, June 2009-April 2010





ANIMALS AND LANDSCAPE SCULPTURE



Outdoor abstract sculptures reside with the land and its residents: trees, grass, flowers, insects, and animals. In turn, artworks influence the behavioral space of animals as they move, play, and pose.

Animals provide a presence contrasting with abstract artworks and, for the sheep visiting *Escaping Flatland*, illustrate reflections and shadows produced by a piece. In photographs of sculpture, the presence of animals provides approximate, informal scaling comparisons with artworks. Sometimes the animals are better at posing in sculpture photographs than humans, especially sheep and Zerlina the dog.



Antoine Durenne, cast iron lion, 19th century, France

Geometric Cutouts 1975 concrete



A hawk visits our sculpture garden, perches on *Tong Bird of Paradise*, and leaves shadowy motion traces upon departure from *Tong Bird*.

Animals have forever served as models and inspirations for art, as in *Tong Bird of Paradise* here and in *Magritte's Smile* at far right.



Magritte's Smile 2009 aluminum casting
length 12 feet or 3.7 meters



SEE NOW . . . WORDS LATER

Suppose this work appeared with a museum description, *A Rare Byzantine Orthodoxy Sacred Cross* (a strong narrative for many). Or the title is *Hommage in Steel for JC* (a reference to Joseph Cornell's collage boxes filled with bird images)? Or the news is out that the artwork sold for 1.2 million euros (now it looks precious indeed)? Or a curator denounced it as a 19th-century fake (oops)? Or that it was made from the bars of torture cells melted down by freed political prisoners? Or that the piece celebrates or ironizes an open-ended implement wrench, an *objet trouvé*? Or how about *The Chicken Goes up the Hill*, a title that makes it impossible to see anything other than an inclined chicken climb?

Instead, give the piece a chance.

Abstract sculptors make objects that generate *unique optical experiences* in the real world. These one-off experiences exist utterly independent of artchat. Abstract artworkers often insist “our only language is vision.”

Seen real by viewers walking around outdoors, abstract works provide a multiplicity of direct and vivid optical experiences: form, scale, color, shadows, volume, airspace, landscape. Direct optical experiences are universal, produced by nature's universal laws that determine how light bounces around 3D artworks. Light from abstract sculptures is focused by the lens of the eye onto the retina. Optic nerves link retinal images to the brain and download optical information at 10 Mb per second, a high-speed connection equivalent to an Ethernet. What then?

Our minds are quick to convert new optical experiences into familiar stories, favored viewpoints, comforting metaphors. No wonder, for how else can we manage optical data flows of 10 Mb per second without familiar categories for filing, without the rage for wanting to conclude?



The rage for wanting to conclude is one of the most deadly and most fruitless manias to befall humanity. Each religion and each philosophy has pretended to have God to itself, to measure the infinite, and to know the recipe for happiness. What arrogance and what nonsense! I see, to the contrary, that the greatest geniuses and the greatest works have never concluded.

Gustave Flaubert

Pre-installed narratives, categories, stale metaphors, reminiscences, and *déformation professionnelle* all interfere with how and what we see. In looking at abstract artworks, once words and story-telling starts, it's hard to see anything else. There's further mischief. The more subtle the object seen and the more precise the distinctions to be made, the more old words deform new seeing. And so Hamlet dominates the seeing of the suggestible Polonius:

Hamlet: Do you see yonder cloud that's almost in shape of a camel?

Polonius: By the mass. And 'tis like a camel indeed.

Hamlet: Methinks it is like a weasel.

Polonius: It is backed like a weasel.

Hamlet: Or like a whale?

Polonius: Very like a whale.

To see with fresh eyes and an open mind requires a deliberate, self-aware act by the observer. Abstract artworks represent themselves and should be first viewed for themselves. When looking at an outdoor abstract piece, concentrate initially on the *unique optical experience* produced by the artwork. See as the artist saw when making the piece. Art is art.

A focus on optical experience does not deny stories, it postpones them. Viewing an artwork may eventually evoke interesting narratives or just tedious artchat: recalling similar art or artists, concocting playful tales, realizing how scrap metal was repurposed into art, making judgments about the artist's intentions or character, or contemplating an artwork's provenance, price, politics.

For a while, then, let the artwork stand on its own. Walk around fast and slow, see intensely from up down sideways close afar above below, enjoy the scenic multiplicity of silhouettes shadows dapples clouds sun zenith earth glowing with the metal. Your only language is vision.

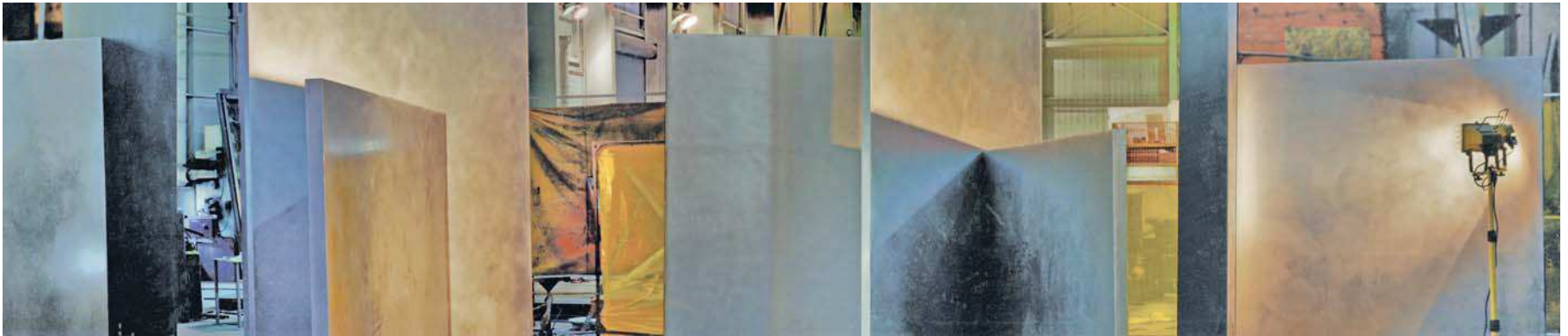


Art is Art (Wrench) 2009 steel
length 8 feet or 2.4 meters

Zenith (Wrench) 2009 aluminum
length 8 feet or 2.4 meters



PRODUCTION AND INSTALLATION: STEEL, GRAVITY, HOPE



Making large artworks requires a factory: big spaces, overhead cranes, heavy equipment, cutting tables, and skilled metal workers who care about sculpture. Artists enjoy working and watching as industrial-scale resources are devoted to producing their pieces. This noisy intense physical metalwork, where art comes alive, differs from the tidy environment of museums and sculpture fields.

Escaping Flatland pieces under construction at Tallix, sculpture foundry in Beacon, New York.

The stainless steel series *Escaping Flatland 1-10* was built at a sculpture foundry; *Rocket Science* at a concrete plant (big indoor and outdoor cranes, flatbed trucks) with scrap metal from a nuclear powerplant; and the large pieces in the series *Tong Bird of Paradise 1-9* were roughed out with cutting torches at a manufacturer of large power equipment and then completed in studio.



The original *Tong Bird of Paradise* (35 inches, 90 cm high) was made by hand, then studied and photographed (here, by an icy pond).

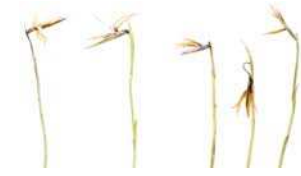


To enlarge *Tong Bird*, silhouette photographs were printed out and artistically redrawn to make a full-scale paper template (40 square feet, 3.7 square meters) of 13 parts for the piece.



The paper template is read by an optical scanner that guides an oxy-acetylene torch that cuts a rusted steel plate to make 5X and 10X linear enlargements of the original small *Tong Bird*.

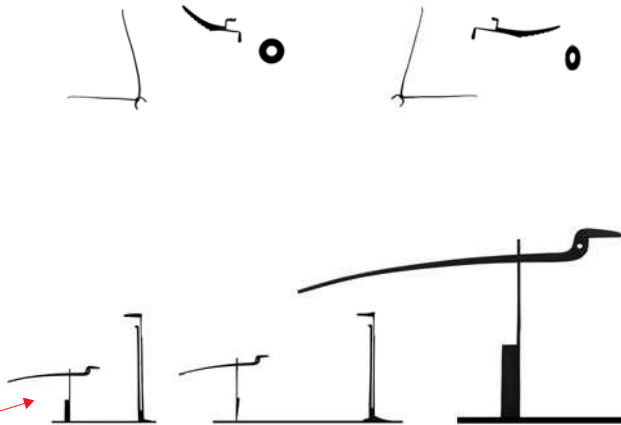
Tong Birds of Paradise began with a pair of forging tongs, which upon disassembly contributed the beautiful curved element for the original small Tong Bird. Then, an old ice saw, a pair of tongs, and a skewed metal circle produced Tong Bird Mobile, a kinetic assemblage that hangs under a dogwood tree:



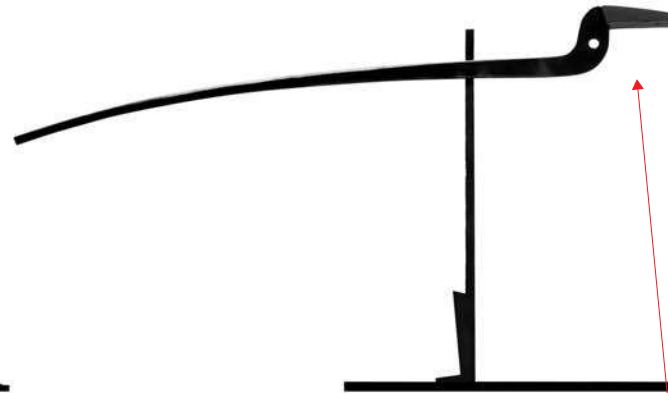
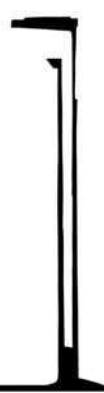
The tall vertical element in Tong Birds of Paradise may resemble a dried Bird of Paradise flower.

Assemblage is different from carving. It is not an attack on things. It is a coming to terms with things. With assemblage or the found object you are caught by a detail or something that strikes your fancy and you adjust, you give in, you cut out, and you put together. It is really a work of love. But there is something else in assemblage, there is the restoration and reparation.

Louise Bourgeois



~ 15 x 10 feet, 5 x 3 meters



~ 30 x 20 feet, 10 x 6 meters



Cutting, grinding, welding, bolting, and assembling turns flat, rough-cut template elements (each 2.75 in, 7cm thick) into a large 3D Tong Bird of Paradise (~ 30 x 20 feet, 10 x 6 meters).

Big Tong Bird was then assembled and installed in a test meadow to study and then adjust its surface (by torch smoking and oiling), visual balance, airspaces, groundline.

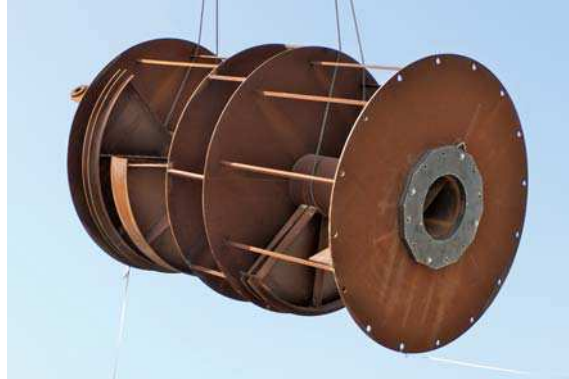
Assorted sculptural elements and their drops for pieces in progress hang around the workshop. A foam model of a fish (Magritte's Smile), awaits shaping, casting, finishing.

Big artworks are big deals to install; outdoor pieces often outweigh humans by hundreds-fold. Steelworkers, riggers, and operators—who do the heavy lifting and putting together—engage in a complex and sometimes dangerous craft. Like sculptors, they move big metal to precise 3D locations, work against gravity, closely attend to rigging operations, and appreciate the sheer physicality of large artworks. Landscape installations are challenging and notable, and everyone involved usually comes away with a good story. Here we see the production and installation of *Rocket Science* (32 feet or 10 meters high; 48,000 lbs or 21,800 kgs of steel).

During an installation, the artist encourages the riggers not to overhandle the piece, evaluates the sculpture's place and posture, and watches with hopeful anticipation as the artwork finally reaches its full public reality. (It is thus a matter of concern if the sculptor's friends at the installation politely say that the new work is "um . . . interesting, very interesting.")

Substantial alterations during installation are difficult and costly. Nonetheless, artists sometimes must make on-site live adjustments, since planning models rarely capture all relevant variables in the high-dimensional space and subtle interactions of sculpture + gravity + light + landscape + time + observer.

During installation, on-site tweaks directed by the artist seek to establish a dynamic balance—in 3-space from assorted viewpoints—between the piece and its land, air, light, shadow. *Dynamic balance* is a mushy verbal concept, but sculptors know it when they see it, and that is what matters.





I am grateful to my colleagues for their good work: Karen Bass, Cynthia Bill, Andrew Conklin, John Fournier, Penny Humphrey, Brian Kelly, Elaine Morse, Jared Ocoma, Kathy Orlando, Andrei Severny, Peter Taylor, Carolyn Williams. Andy Conklin, master metal worker, constructed many of these pieces; several were built with superb craft at Tallix and Polich Tallix in New York; the excellent staff at United Concrete in Yalesville, Connecticut—Jon Gavin, Dennis Brouillard, Mike Nitkowski, Bruce Woronoff—engineered, built, and installed many large pieces; and both John Hilzingers at Heavyweight, Inc found beautiful scrap metal from a nuclear powerplant. I am also deeply grateful for the advice and encouragement of Harry Philbrick and Richard Klein at The Aldrich Contemporary Art Museum, who guided my first major sculpture show, at the Aldrich June 13, 2009 to January 17, 2010. Photographs by Andrei Severny, Philip Greenspun, Dmitry Krasny, and the artist.

EDWARD TUFTE July 2010

COVER *Peanuts* ©1960 Peanuts Worldwide LLC. Dist. by Universal Uclick. Reprinted with permission. All rights reserved.

PAGE 4 Ursula von Rydingsvard, as quoted in “A Conversation with Ursula von Rydingsvard: Objects of Presence,” *Works & Conversations* 8, interview by Richard Whittaker, 2003.

PAGE 8 David Smith, as quoted in *David Smith by David Smith* (1968), 133.

PAGE 15 Gyorgy Kepes, *The Language of Vision* (1948). Henry Moore, quoted in *Sculpture and Drawings by Henry Moore* (1951), 4.

PAGE 22 Richard Serra, *Writings Interviews* (1994), 170-171.

PAGE 30 Kristin Koch, Judith McLean, Ronen Segev, Michael A. Freed, Michael J. Berry, Vijay Balasubramanian, Peter Sterling, “How Much the Eye Tells the Brain,” *Current Biology* 16 (July 25, 2006), 1428-1434. Gustave Flaubert, *Correspondance* (Paris, 1929), volume v-III, translated by Dawn Finley.

PAGE 33 Louise Bourgeois, *Deconstruction of the Father, Reconstruction of the Father: Writings and Interviews 1923-1997* (1998), 143.

Copyright ©2013 by Edward Rolf Tuftte Published by Graphics Press LLC
P.O. Box 430, Cheshire, Connecticut 06410 USA WWW.TUFTE.COM

All rights to text and photographs are reserved by Edward Rolf Tuftte. This work may not be copied, reproduced, or translated in whole or in part without written permission of the publisher, except for brief excerpts in reviews or scholarly analysis. Use with any form of information storage and retrieval, electronic adaptation or whatever, computer software, or by similar or dissimilar methods now known or developed in the future is strictly forbidden without written permission of the publisher and the other copyright holders.

Edward Tufte has had solo shows of sculptures and prints at Artists Space in New York, the Architecture+Design Museum in Los Angeles, and a major sculpture show at the Aldrich Contemporary Art Museum in 2009-2010. Since 1999, he has completed 50 large-scale outdoor pieces, 120 table pieces, and many steel engravings and digital prints. ET's first abstract artwork was drawn, at age 6 years old, in ink with mechanical drafting pens:



Edward Tufte wrote, designed, and self-published 4 books on analytical design: *The Visual Display of Quantitative Information* (1983, 2001), *Envisioning Information* (1990), *Visual Explanations* (1997), and *Beautiful Evidence* (2006). *The New York Times* described ET as the “da Vinci of data” and *Business Week* as the “Galileo of graphics.” These books have received 30 awards for content and design and have 2 million copies in print. He has received 7 honorary degrees, and is a Fellow of the American Academy of Arts and Sciences, the Guggenheim Foundation, the Center for Advanced Study in the Behavioral Sciences, the Society for Technical Communication, and the American Statistical Association. He served as a Professor at Princeton University and Yale University for 32 years.

ET's sculpture fields (visits by appointment) are on 234 acres in Cheshire and Woodbury, Connecticut. Photographs, videos, and accounts of the artworks are at www.tufte.com.

