Menu

Table of Contents

Measures of Association and Regression Analysis

"How did I get into this business? Well, I couldn't understand regression and correlation in college, so I settled for this instead."

Lyman Ott, Richard F. Larson, & William Mendenhall, Statistics
(Boston: Duxbury Press), 1987: 430.

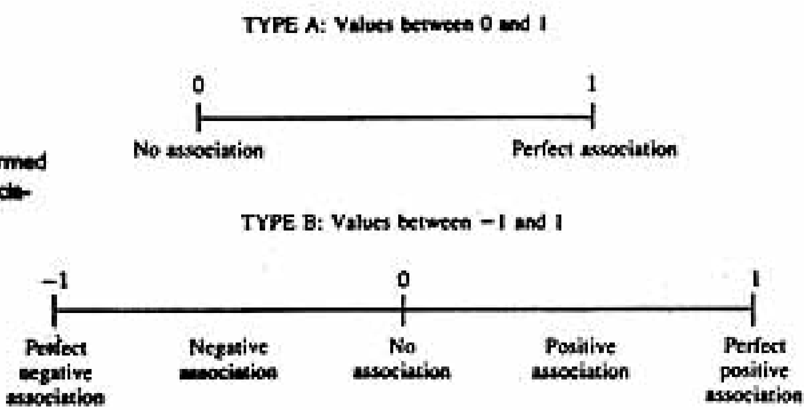
From the Literature

Howard Schuman and his colleagues reported on their investigation of the relationship between effort and grades in college (*Social Forces*, June 1985). The researchers recognized two widely held assumptions: (1) hard work along socially prescribed lines produces rewards and (2) the relationship between hard work and rewards is valid even for most researchers who emphasize the importance of other variables, such as natural ability and sheer luck.

Schuman and colleagues had as the overall focus of their study the test of the popular maxim "Genius is one percent inspiration and ninety-nine percent perspiration." To test this maxim, as well as a number of specific hypotheses, they drew upon the responses of a random sample of 522 students attending a major midwestern university. The interview questionnaire tapped such variables as hours studied, grade point average (GPA), and SAT scores.

GPA and SAT scores were validated later by obtaining access to official student records. The variable "hours studied" was divided into two components: hours studied during the week and hours studied on weekends. These two components were moderately related (r = .47). Small positive associations were found individually between GPA and hours studied (r = .11), as well as between GPA and the percentage of classes attended (r = .28). Taken as a group, classes attended, hours studied, and total SAT scores were found to provide a meaningful prediction of GPA.

The investigators also reported that performance on early tests was highly correlated with performance on later tests, and that later tests were even more highly correlated with the final grades earned in a class (r varied between .81 and .88). Other patterns were also reported. They included the finding that the total number of hours typically spent studying was related to the number of hours studied yesterday (r = .74), scores on the first test (r = .82), scores on the second test (r = .84), and scores on the third test (r = .86). Finally, SAT scores, by themselves, were related to GPA (r = .44).



Two types of normed measures of association

Surprising Correlations

Sometimes correlation coefficients only tell us what we already know by confirming the obvious. Just as often, however, unsuspected relations emerge from the data.

Take the relation between the effectiveness of teachers and the subjective ratings students give them. It is only reasonable to suspect that college students know a good teacher when they see one and, when asked to grade the performances of their instructors, assign the best grades to the best teachers. One recent study suggests, however, that just the opposite may be true. Roden and Rodin approached this issue by developing a highly objective measure of the amount of material learned from the instructor in an undergraduate calculus course.[1] They calculated the mean amount learned by each of twelve classes and then determined the mean rating each class assigned its teacher. The scatter plot in Figure 6.9 shows the surprising results: A negative correlation between subjective ratings of teachers and amount learned is clearly indicated by the slope of scatter-plot points, and the computed correlation coefficient for these data was $r = -0.75$. Rodin and Rodin appropriately point out that any explanation for the negative correlation offered at this point would be speculative, but they do interpret the results to indicate that "students are less than perfect judges of teaching effectiveness if the latter is measured by how much they have learned."[2]

Or, consider the relation that surely must exist between students' backgrounds in mathematics and the grades they receive in statistics courses. Undoubtedly, the student entering a college sophomore-level statistics course having already taken calculus and advanced algebra is better off than a student who has had only elementary algebra. But are the highest statistics grades received by those with the more extensive math backgrounds? The answer is, "That's an empirical question." This means that, strong as our intuitions on the matter may be, the only way to find out for sure is to examine the evidence.

Dr. Leonard Giambra addressed this question by examining the relation among mathematics backgrounds, overall grade-point averages, and grades received by 201 students in his statistics classes.[3] Surprisingly, he found no apparent relation between statistics grades and math backgrounds--students who had taken calculus did not, for example, receive a disproportionate percentage of the A's and B's. There was, however, some correlation (0.25) between statistics grades and overall grade-point averages, suggesting to Dr. Giambra that a student's grade in statistics depends more on his overall ability than on his math background.

The results from these two studies may or may not apply to situations beyond the immediate ones in which they were obtained. But they effectively illustrate the utility of correlational analysis in distinguishing between objective reality and cherished suppositions.

1M. Rodin and B. Rodin, "Student Evaluation of Teachers," Science 177 (1972): 1164-1166.
2Ibid., p. 1166.
3L. Giambra, "Mathematical Background and Grade-Point Average as Predictors of Course Grade in an Undergraduate Behavioral Statistics Course," American Psychologist 25 (1970): 366-367.

Understanding and Using Statistics-Basic Concepts, Marty J. Schmidt, pp. 175-177.

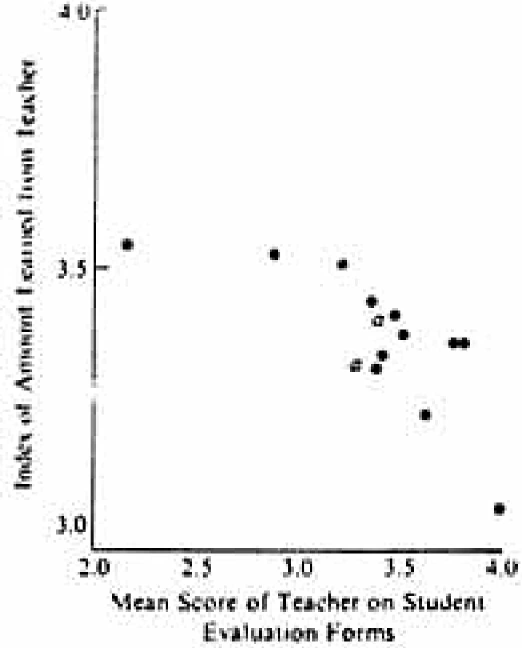


Figure 6.9 Relation between objective and subjective criteria of good teaching ($r = -0.75$). The points labeled *a* are for two sections taught by the same instructor.

Source: From M. J. Rodin and B. Rodin, "Student Evaluation of Teachers," *Science*, 177 (September 1972): 1164–1166. Reprinted by permission from Dr. Miriam J. Rodin and Dr. Burton Rodin and the American Association for the Advancement of Science. Copyright 1972 by the American Association for the Advancement of Science.

Measures of Association

There are many ways to measure association and each one may be interpreted differently. These comments amplify the discussion in the text and should help in interpreting these measures.

The measures most often used in sociology are *gamma, tau, rho, lambda,* and r (Pearson correlation coefficient). Most measures of association, and all of the ones just mentioned, have values between 0 and 1. Values close to 0 indicate a low relationship, those near 1 indicate a high relationship. In other words, the higher the value the greater the relationship between any two variables.

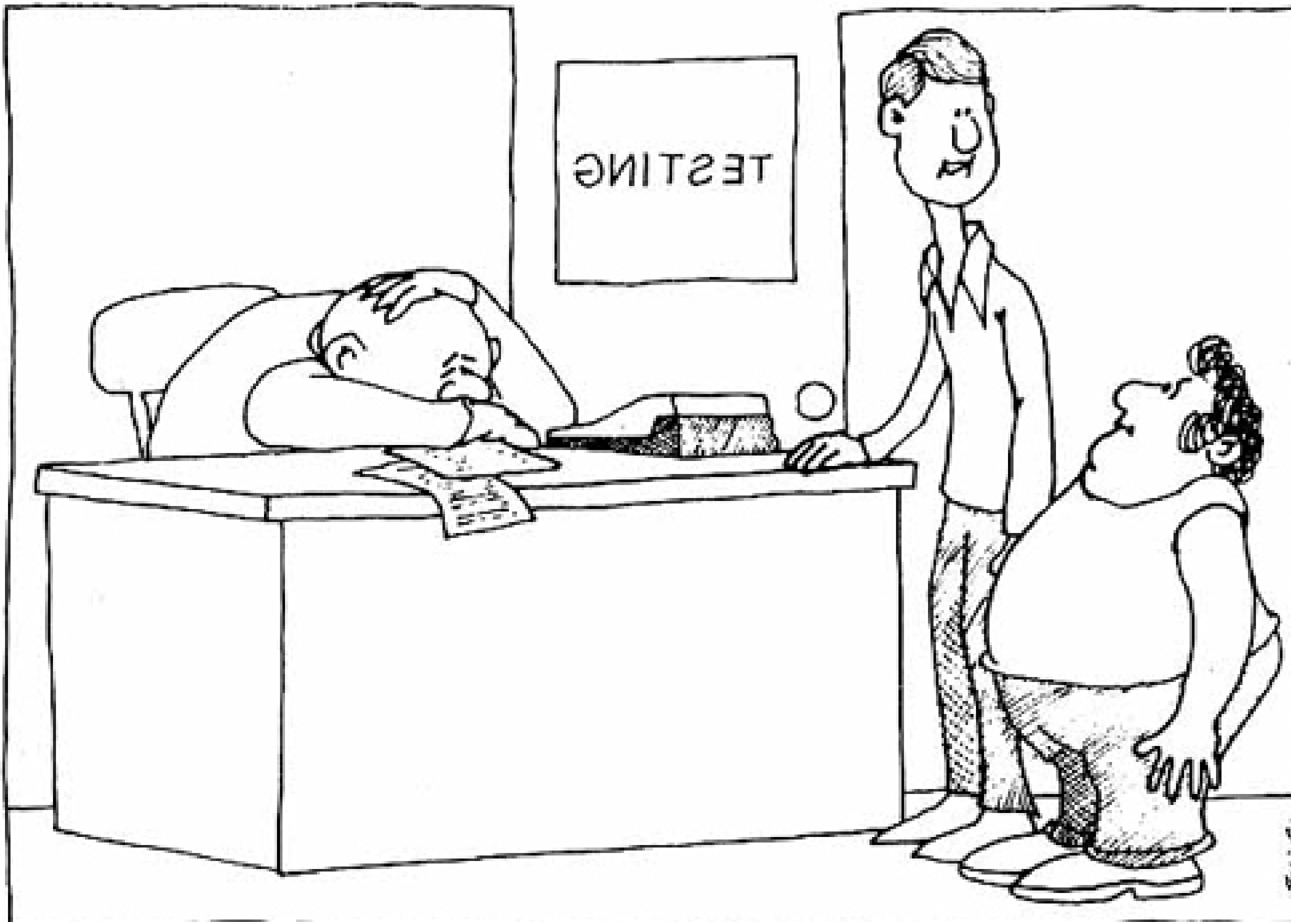
The direction of the relationship is also included in these statistics. Direction can be positive (direct) or negative (inverse). A positive relation obtains when one variable increases as the other variable increases, or when one decreases as the other decreases. The two variables move or change in the same direction. A perfect positive relationship equals +1.00. On the other hand, a negative relation obtains when one variable increases as the other decreases. A perfect negative relationship equals −1.00.

There is no one simple interpretation for the various statistics. However, for the most commonly used measure, r, the simplest interpretation is in terms of r^2, which is the variance explained. If our $r = .70$, $r^2 = .49$, which is the proportion of the variance explained by one variable on the other. With other measures, it is necessary to rely on the magnitude of the relationship in order to interpret it; the closer to 1 the particular measure is, the greater the relationship between them.

In order to make easier a decision about the correct interpretation of these measures (percentages, rates, and measures of association), sociologists frequently report a "P value." Generally at the bottom of a table you will see $P = .05$ (or some other value), or a statement like, "correlation significant at 0.05 level." This statistic is telling you that the correlation is not due to chance. Put another way, if the study were done over again, you would find approximately the same correlation.

In the social sciences, if a relation could occur at the .05 level, or .01 or .001, it is considered statistically significant. To say that a finding is statistically significant is only to say that the finding is very unlikely to be due to chance. To say that this is not a chance finding, however, is not to say what it is. And it should not be confused with being theoretically or practically signifitant. However, it does mean that whatever was found is likely what is there, that is, if you do the study again, you will likely find the same thing.

Robert Hagedorn, (General Editor), Sociology, Dubuque, Iowa: William C. Brown, 1983: 57.



"He says we've ruined his positive association between height and weight."

Correlation Versus Cause and Effect

Darrell Huff, in How to Lie with Statistics,[1] reports there is a positive correlation between the ages of women and the angles of their feet in walking. Younger women tend to point their toes straight ahead; older women tend to walk with toes pointed out. Is it possible that the aging process causes women's feet to turn out? When one assumes, because two variables are correlated, that changes in one cause changes in the other, one has a post hoc error. The above relationship no more proves that aging causes feet to turn out than it proves the opposite--turning the feet out causes an increase in age.[2]

There are times, of course, when high correlations appear and no one is tempted to make the post hoc error. Reportedly, there are high correlations between (1) the sizes of schoolboys' feet and the quality of their handwriting;[3] (2) the number of storks' nests and the number of human births in northwest Europe;[4] and (3) the salaries of Presbyterian ministers in Massachusetts and the price of rum in Havana.[5] In some cases we can suggest a possible third factor that might account for the relation, but tend to be older than children with smaller feet, and handwriting probably improves with age. The number of storks' nests in Europe is related to the number of chimneys, which is related to the number of houses, which is related to the number of human births. The reason for relation 3 is anybody's guess.

There are other times, however, when we suspect that correlated variables might indeed represent a cause-and-effect relationship, but we cannot jump to that conclusion strictly on the basis of our correlational information. If, for instance, a medical study found that people who drink more coffee tend to have more coronary heart disease problems than people who drink less coffee, could we conclude that coffee drinking causes heart problems? It is possible that people who drink more coffee have, on the average, more sedentary jobs and thus get less daily exercise. Not until we can rule out all other possible causes can we begin to make cause-and-effect arguments based on correlational data. This is the problem faced by the research specialist in epidemiology (a science that attempts to understand and control disease in populations). Dr. Paul Milvey, a biophysicist who specializes in epidemiological problems, discusses some of the problems involved in demonstrating a causal link between physical exercise and risk of coronary heart disease (CHD):

> Most of the several hundred studies that deal with the relationship of physical activity to CHD or mortality from all diseases show the general picture of those people who are physically active tending to have a lower incidence of disease at any particular age than those who are not.
>
> The scientific and medical field pursuing these studies is called epidemiology. Epidemiologists seek to discover through statistical studies of man and his environment the cause or causes of pathology and disease. Whether the problem is cholera in 19th century England (solved), the Legionnaires' disease in Philadelphia last year (solved), the primary cause of lung cancer (solved), bladder cancer (unsolved) or CHD (unsolved), epidemiologists seek to demonstrate a cause-and-effect relationship between two entities.
>
> The criteria for demonstrating this relationship are quite specific and are difficult to satisfy. Sophisticated mathematical techniques often are employed so the epidemiologist can achieve valid conclusions. It's a difficult discipline because, while it's very easy to prove "association," it is infinitely more difficult to demonstrate "cause and effect."

For example, in the 1960s (Michael) Yudkin studied the relationship between dietary sugar and CHD and its final manifestation, the heart attack. He took very careful dietary histories of three groups of hospital patients: (1) patients hospitalized for any one of a large variety of reasons (broken legs to appendectomies); (2) patients suffering from CHD; and (3) patients who had had myocardial infarctions (heart attacks).

The amount of sugar each group consumed over the months and years prior to hospitalization was shown to be least for the "control" patients with the variety of diseases unrelated to their hearts and arteries, intermediate for the CHD patients and highest for the heart attack patients.

This study seemed strongly to indicate (nothing is ever quite proven in science) that sugar consumption caused or perhaps was one of several causing factors in the development of CHD and its final manifestation, the heart attack. But several investigators questioned this.

In the 1970s, two groups, working independently, showed that there was a better association or correlation between the development of this disease and the amount of the patients' smoking. And there was an even better correlation--and excellent correlation--between smoking and sugar consumption. (Those who smoked more also consumed more sugar.)

From a scientific and medical point of view, there is no obvious reason why sugar and CHD should have a causal relationship (fats, especially saturated fats, are quite probably related to CHD, but not sugar). But one can suggest any of a number of good, medically sound explanations or mechanisms by which smoking might cause CHD. And there is statistical evidence that also shows this. So we were fooled for a number of years by the earlier study which showed a correlation or association between CHD and sugar consumption, when in fact no causal relationship existed.

In this case, the epidemiologist calls sugar a "confounding variable." It goes along with the real causal variable. It's absent when the real cause is absent, it's there when the real cause is there, and it's a damned confusing problem to handle in all epidemiological studies.

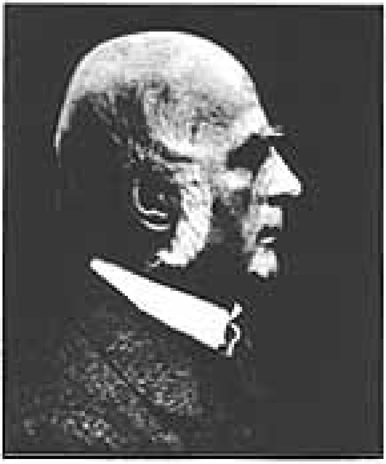
Why are sugar and smoking consumption correlated? We can only speculate. We do know that the lower socio-economic groups in our country eat more sugar (junk food) than the higher socio-economic classes. Although I haven't checked it out, they almost certainly smoke more as well. So both smoking and sugar consumption may, in this light, be seen to be "caused" by one's socio-economic status in life.

We also know the CHD is higher in the less affluent classes. So perhaps the cause of CHD should be looked for in some common aspect of the environment or life style of the less-advantaged socio-economic classes: Do they go for checkups less frequently, do they smoke and consume fats or salt in larger amounts?

It's very difficult for the epidemiologist to pinpoint the environmental causes. . . .[6]

[1]Darrell Huff, How to Lie with Statistics (New York: W.W. Norton, 1954), Chapter 8.
[2]Huff suggests that older women were raised during a time when toeing out was encouraged; young women today are encouraged to walk with a different posture.
[3]W.A. Wallis and H.V. Roberts, The Nature of Statistics (New York: Free Press, 1962), p. 108.
[4]Ibid., p. 108.
[5]Huff., How to Lie with Statistics, p. 90.
[6]From P. Milvey, "Getting to the Heart," Runner's World, 12 (April 1977): 27-31.

The Bettmann Archive.



SIR FRANCIS GALTON

The least-squares method will happily fit a straight line to any two-variable data. It is an old method, going back to the French mathematician Legendre in about 1805. Legendre invented least squares for use on data from astronomy and surveying. It was Sir Francis Galton (1822–1911), however, who turned "regression" into a general method for understanding relationships. He even invented the word.

Galton was one of the last gentleman scientists, an upper-class Englishman who studied medicine at Cambridge and explored Africa before turning to the study of heredity. He was well connected here also: Charles Darwin, who published *The Origin of Species* in 1859, was his cousin.

Galton was full of ideas but was no mathematician. He didn't even use least squares, preferring to avoid unpleasant computations. But Galton's ideas led eventually to the machinery for inference about regression that we will meet in this chapter. He asked: If people's heights are distributed normally in every generation, and height is inherited, what is the relationship between generations? He discovered a straight-line relationship between the heights of parent and child and found that tall parents tended to have children who were taller than average but less tall than their parents. He called this "regression toward mediocrity." Galton went further: he described inheritance by a straight-line relationship with responses y that have a normal distribution about the line for every fixed input x. This is the model for regression we use in this chapter.

It Tastes Good, But Is It Good for You?

Sometimes regression lines are useful for purposes other than for predicting future events from past events. The coefficients that define the line--slope and intercept--are sometimes valuable in their own right for analyzing characteristics of a linear relation between two variables.

Consider, for instance, a problem faced by two General Food Corporation researchers, Elisabeth Street and Mavis Carroll. They were involved in an attempt to develop an "easy-to-prepare, nutritious, on-the-run meal," code named H.[1] In the laboratory it is quite easy to ensure that such concoctions contain certain nutrients, but there is little a prior insurance that humans will metabolize those nutrients efficiently; there is even less assurance that they will find them palatable. Such factors must be determined empirically by testing the finished product on living subjects.

A rigorously controlled test using human subjects indicated that, indeed, H was as tasty as another product C (so called because it has a casein base), which was already on the market. The researchers were also interested, however, in finding whether or not the protein content of H would be metabolized efficiently under conditions of actual use. They approached this question in another carefully controlled study using rats as subjects. Rats--rather than humans--were used in this study because the metabolic processes involved in protein utilization are fairly similar in rats and humans and because the relative efficiency of that utilization is more clearly and quickly reflected in the animals' body weights.

Thirty rats were randomly divided into three groups of ten each; each group contained animals of comparable weights, that is, all groups had equal-age light, medium, and heavy rats.[2] All animals were individually weighed before the experiment, and then each group was given a different diet for 28 days. One group was fed only on a liquid version of H, another group received only a solid version of H, and the third (control) group was fed on casein. Each of the three diets contained about 9 percent protein by weight. During this experimental period the animals were allowed to eat as much as they wished of the designated food, and the food intake of each animal was carefully monitored.

At the end of the 28-day feeding period, the researchers had two measures on each animal: body weight gain (in grams) over the 28 days, and the 28-day protein intake. These data constitute paired observations on two variables for each rat and are displayed in the scatter-plot diagram shown in Figure 7.8. Each dot or cross thus represents data from one animal.

Figure 7.8 shows clearly that animals on liquid and solid H took in more protein and gained more weight than animals on the casein diet. The researchers reasoned, however, that more analysis of these data was needed in order to answer the question of whether or not H protein was used more efficiently than casein protein. It is possible, after all, that H was simply more tasty than casein. Casual inspection of Figure 7.8 does not indicate whether or not a gram of H protein can be expected to produce a greater weight gain than a gram of casein protein. So, a separate regression line was computed for each group's paired observations. These are shown (without data points) in Figure 7.9.

Notice that the H diets produced steeper regression lines than did the casein diet. The calculated slopes, b, of the liquid and solid H groups were respectively, 3.72 and 3.66, while b for the casein group was 2.91. So, the researchers concluded, differences in weight gain between H and casein groups were not due entirely to differences in protein intake. The differences in b values indicated that "For a given increase in protein intake, the H diets resulted in a greater increase in weight gain than did the casein diet."[3]

[1]E. Street and M.B. Carroll, "Preliminary Evaluation of a New Food Product," In J. Tanur et al. (eds.), Statistics: A Güide to the Unknown (San Francisco: Holden-Day, 1972), pp. 220-228.
[2]The procedure used here was similar to the matched pairs experimental design. Here, matched trios of rats were arranged before the experiment; one member of each trio was randomly assigned to each group.
[3]Street and Carroll, op. cit., p. 224.

Understanding and Using Statistics-Basic Concepts, Marty J. Schmidt, pp. 198-201.

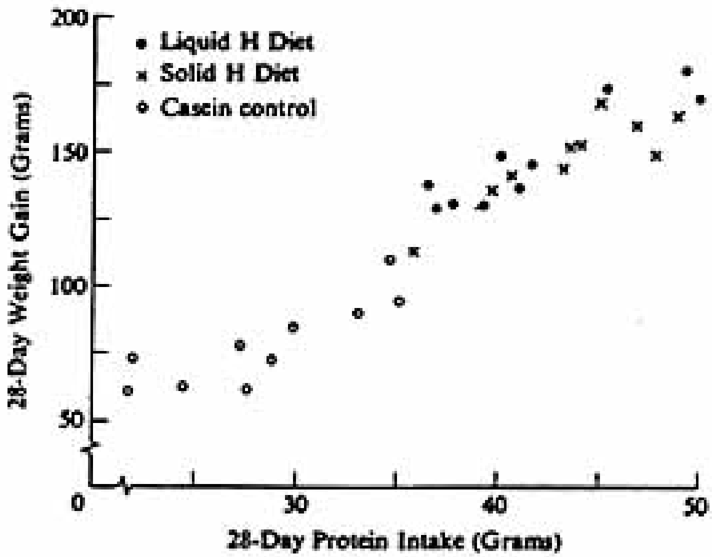


Figure 7.8 Relationship of 28-day protein intake and weight gain in young male rats.

Source: Figures 7.8 and 7.9 are taken from E. Street and M. B. Carroll, "Preliminary Evaluation of a New Food Product," in J. Tanur et al. (eds.), *Statistics: A Guide to the Unknown* (San Francisco: Holden-Day, 1972), p. 223. Reprinted by permission.

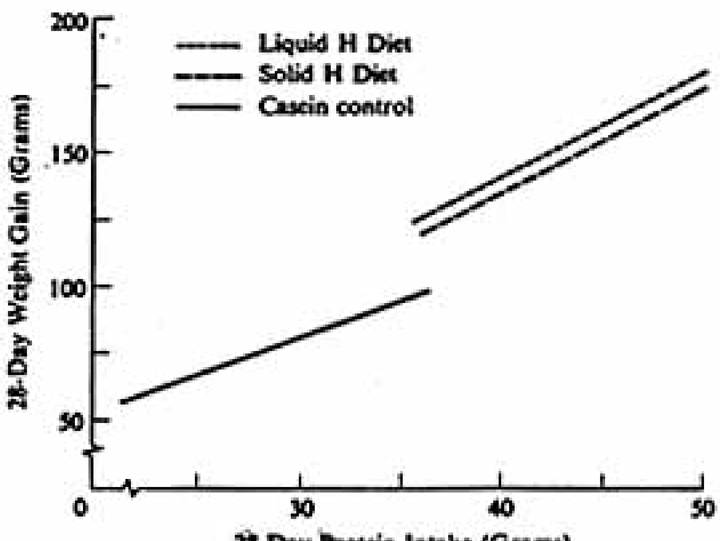


Figure 7.9 Estimated regression of 28-day weight gain on protein intake for young male rats.

"It was a numbers explosion."

Reproduced by permission of Bob Schochet.

SCRUTINIZING POPULAR REPORTS OF SOCIAL SCIENCE RESEARCH

We are being bombarded daily with such a mass of new information that it is difficult to process it adequately. It is increasingly important to become a critical, selective, and informed consumer of information. Discussed below are several means for better evaluating reports on social science research that you may encounter in the media.

Maintain a Skeptical Attitude

Be skeptical, because the media have a tendency to sensationalize and distort. For example, the media may report that a university researcher spent $500,000 to find out that love keeps families together when, in fact, this was only one small aspect of the larger research project. Moreover, chances are the media have oversimplified even this part of the researcher's conclusions.

Consider the Source of Information

It is important to know, for example, whether a study on the relationship between cancer and smoking has been sponsored by the tobacco industry or by the American Cancer Society. On a 1985 "60 Minutes" segment, a representative of a tobacco company denied the existence of any research linking throat and mouth cancer with dipping snuff. A medical researcher contended that putting a "pinch between your cheek and gum" has, in the long run, led to cancer in humans. Whom do you believe? At the very least you want to know the background of the source of information before making a judgment about scientific conclusions.

Determine Whether a Control Group Has Been Used

Knowing whether a control group has been used in the research may be important. For instance, increases in self-esteem and physical energy may be reported in a study of participants in a meditation program. Was this because of the respect and attention they were given during the training period or because of the meditation techniques themselves? Or a study may report that the productivity of a group of workers in an office increased dramatically because the workers were allowed to participate in work-related decisions. Was the productivity increase due to the employees' being involved in something new and exciting or because of the participation in decision making itself? Without one or more control groups, you cannot be certain of what caused the changes in the meditation participants or in the office workers.

Do Not Mistake Correlation for Causation

A correlation between two variables does not necessarily mean that one caused the other. For example, at one time the percentage of Americans who smoked was increasing at the same time life expectancy was increasing. Did this mean that smoking caused people to live longer? Actually, a third factor—improved health care—accounts for the increased life expectancy. Do not assume that two events are related just because they occur together.

STATISTICAL DESIGN AND THE EQUINE QUADRUPED

The genesis of modern-day statistics seems to be accounted for, in great part, by the efforts of an English gentleman named R.A. Fisher. Much of Fisher's research was done during the twenties and before, with the beginning of his scholarly publications in the early 1930's. Fisher's work was with agricultural experimentation, which ultimately led to blocking theory and split-plot statistical design. It is his agricultural research that was most interesting to the author, and the remainder of the paper will be devoted to a systematic and logical analysis of the genesis of statistical design as it relates to the equine quadruped.

Fisher's early work was done with garden plots and was initially as much an interest in flowers and truck products as it was with statistics, but being a mathematician he became interested in statistical design as a means of recording, controlling, and analyzing the results of his many and varied experiments.

In addition to experimental design as a means of controlling variables, Fisher exercised other controls with regard to plant growth. One of the major variables was the use of fertilizer, about which the author will take the liberty of making some assumptions.

Much of Fisher's experimentation took place during the twenties, perhaps some before then, and on in to the thirties. Since this period of time was prior to the widespread use of synthetic and chemically contrived fertilizers, the author will assume that natural fertilizers were still in wide use and that, further, Mr. Fisher was an ardent user of said natural fertilizers as a means of achieving the "big" effect on the experimental variable.

Further, that one of the major animals of England has, and will always be, the equine quadruped. Said quadruped had long been a mainstay in English society and, yea, even government when we consider the effect of the horse on extending the power and stature of the medieval knight in defense of whatever ruler. Since the ubiquitous quadruped was so prevalent, approached only in numbers perhaps by the lowly bovine quadruped and the chicken, the author will assume that the majority of natural fertilizer was produced by the equine quadruped. That being the case, Mr. Fisher must have been an ardent user of the fecal matter of the equine quadruped as an experimental variable; there seems to be little current question on this point.

What is truly amazing, in the opinion of the author, is that we have come so far in the field of chemical fertilizers only to find the field of statistics still grossly encumbered and shot through and through with horseshit.

SOURCE: <u>Harper's Magazine</u>, January, 1974, p. 5.