San Diego Law Review

Volume 21
Issue 4 *1984*

Article 3

7-1-1984

A Common Sense Approach to Understanding Statistical Evidence

David W. Barnes

A Common Sense Approach to Understanding Statistical Evidence†

DAVID W. BARNES*

This article presents a straightforward and intuitive method for understanding and interpreting statistical evidence submitted to courts as proof of factual issues. The goal is to overcome the reader's fear and loathing of statistics by relating all statistical methods to the concepts of numerical differences between numbers and similarities or correspondences between numbers. Throughout the article, the terminology is in conversational rather than technical English and actual rather than hypothetical cases are used to illustrate and explain statistical tools which gradually increase in complexity as the reader progresses. Cases are drawn from a wide variety of substantive law areas such as civil rights, employment discrimination, contracts, environmental law, energy law, constitutional law, deceptive advertising, and highway traffic safety. The discussion begins with the concept of subtraction and proceeds through percentages and correlations to regression analysis. Using the statistical concept of a standard deviation, which is explained in intuitive terms, statistical evidence of all degrees of complexity is described as a mechanism for ascertaining whether an absolute magnitude or measurable effect is big enough to be legally significant.

---

† Since the initial drafting of this article, the topic has been expanded into a treatise on the use of statistical evidence in American courts forthcoming from Little, Brown and Co. in 1985. Authors of the treatise are David W. Barnes, John M. Conley, and David W. Peterson. The author of this article is wholly responsible for any oversimplifications and inaccuracies appearing herein. He wishes, however, to express his great appreciation to his co-authors of the treatise for their critical comments and to Carol Bronson and Lisa Ivy Cohen as well, for their continuing contributions to his research.

* A.B., Dartmouth College, 1972; M.A., Virginia Polytechnic Institute and State University, 1976; J.D., University of Pennsylvania College of Law, 1979; Ph.D. (economics), Virginia Polytechnic Institute and State University, 1980. Mr. Barnes is Director of the Center for Interdisciplinary Legal Studies and an Associate Professor of Law and Economics at the Syracuse University College of Law.

INTRODUCTION

A baffling array of statistical tests is offered in courts and administrative tribunals to prove competing factual positions. These competing methods appear to be unrelated to one another and equally valid or invalid, depending on the perspective and predilections of the uninitiated observer. Full utilization of quantitative evidence as a tool of advocacy requires an appreciation of available options. One can drive twelve penny nails with a tack hammer but a two-pound driver is demonstrably more powerful. This note describes the utility of a full toolbox by reference to statistical tools of increasing complexity. The purpose is not to teach how to calculate statistics or manipulate numbers. There are innumerable texts designed to teach statistical theory and mathematical computations. There is even one designed to teach these tools to law students,[1] and there are several for practicing attorneys.[2] This note develops a new, simple, intuitive basis for understanding what these statistical tests mean and where they are useful. It starts with the concept of subtraction and progresses to a basic understanding of the utility and interpretation of multiple regression.

DIFFERENCES AND CORRELATIONS

Differences

Basic subtraction is used to determine how different one number is from another. In *Castaneda v. Partida*[3] the criminal defendant alleged that Mexican-Americans were underrepresented on grand juries in Hidalgo County, Texas, where he was convicted of burglary with intent to rape. His counsel argued that if Mexican-Americans were represented on juries with the same frequency that they appeared in the population as a whole there would have been 688 Mexican-American grand jurors over the past eleven years instead of the 339 actually observed.[4] The simple process of subtracting the observed number from the expected number of Mexican-American jurors yields a disparity or difference of 349 jurors. This disparity summarizes a relevant characteristic of the expected and the observed number, that is, the difference between them.

For purposes of legal proof, subtraction of whole numbers is conceptually flawed because it is insensitive to the absolute magnitudes

1. D. BARNES, *Statistics as Proof: Fundamentals of Quantitative Evidence*, (1983).

2. *See generally* D. BARNES, J. CONLEY, AND D. PETERSON, *Statistical Evidence in Litigation*, forthcoming in 1985; CURTIS, *Statistical Concepts for Attorneys: A Reference Guide* 1983.

3. 430 U.S. 482 (1977).

4. *Id.* at 487 n.7.

of the numbers involved. Subtracting a hypothetical 17,339 observed jurors from 17,688 expected also shows a difference of 349 jurors but this shortfall of expected Mexican-American jurors seems much less serious given the large total number of jurors selected.

Percentages are adopted as a means of eliminating the distorting effects of absolute magnitudes. In Hidalgo County, 79.1% of the population was Mexican-American but only 39% of the people called for jury duty were Mexican-Americans,[5] a difference of 40.1 percentage points. With the larger hypothetical numbers, the difference is only 1.6 percentage points[6] thus confirming our intuitive impression that the hypothetical disparity is not as serious as the actual example from *Castaneda.*

Courts' interpretations of the difference between percentages appear to be consistent with this impression. In *United States v. Goff*,[7] Blacks and food stamp recipients alleged that they were underrepresented on the voter registration list from which federal grand jurors were drawn. The disparities between expected and observed percentages were 5.27 percentage points for Blacks and 6.17 percentage points for food stamp recipients. The *Castaneda* court found that 40.1 percentage points was a significant disparity evidencing discrimination[8] while the *Goff* court found 6.17 percentage points insufficient to show substantial underrepresentation of food stamp recipients.[9] These findings are understandable; 6.17 percentage points is a considerably smaller disparity than 40.1 percentage points.

Sometimes courts find very small percentages to be indicative of serious injustices. For example, in *Board of Education v. Califano*,[10] only 0.9% Black teachers were employed at a school where, had teachers been assigned in a racially neutral fashion, one would expect 5.1% of the teachers to be Black. This underrepresentation of 4.2 percentage points (5.1% minus .9%) was thought by the court to be substantial and evidence of disparate treatment of Whites and Blacks.[11] Why is 6.17 percentage points insufficient in one case but 4.2 percentage points sufficient in another to show discrimination? There must be something more involved here than simple differences

5. *Id.* at 486-87.
6. 17,688 = .791 x 22,362 and 17,339/22,362 = .775. The percentage point difference is .791 - .775 = .016 or 1.6 percentage points.
7. 509 F.2d 825 (5th Cir. 1975), *cert. denied*, 423 U.S. 857 (1975).
8. 430 U.S. at 496.
9. 509 F.2d at 827.
10. 584 F.2d 576 (2d Cir. 1978), *aff'd*, 444 U.S. 130 (1979).
11. *Id.* at 589.

between either absolute numbers or percentages. Fancy statistical methods combined with knowledge of the relevant law are necessary to resolve this enigma.

Instead of two-group cases, Black/non-Black in *Board of Education* or Mexican-American/non-Mexican-American in *Castaneda*, consider a case where there are numerous minority groups, each of which is allegedly subject to discriminatory treatment. In such a case, a summary of the overall discriminatory impact might be desired given the dissimilar treatment of each group. *Inmates of the Nebraska Penal and Correctional Complex v. Greenholtz*,[12] in which there were four different racial groups, is just such a case. Plaintiffs were a class of Native American and Mexican-American inmates of a Nebraska prison. They charged that the defendants, members of the Nebraska Board of Parole, denied discretionary parole to class members on racially and ethnically discriminatory grounds. Statistics showed that 59.3% of all persons eligible were granted release by discretionary parole. While 60.7% of Whites and 63.0% of Blacks were paroled, only 40.7% of Native Americans and 27.8% of Mexican-Americans were paroled.[13]

A similar problem arose in *Chance v. Board of Examiners and Board of Education of New York*,[14] a case in which candidates seeking licenses for permanent appointment to supervisory positions in the City of New York School System alleged unlawful differential treatment of Caucasians, Blacks, and Puerto Ricans. In *Inmates* and *Chance*, the difficulty is not in calculating the differences for each group, but in determining whether the overall disparity is sufficient to indicate unlawful discrimination among the various categories. Increasing the number of groups receiving different treatment complicates the intuitive comparison.

Determining the legal sufficiency, or, as it is referred to in statistics, *significance*, of a difference may also arise when comparing averages from two or more groups. In *In re Forte-Fairbairn, Inc.*,[15] a contract dispute arose over the identity of certain fibers invoiced as baby llama fibers. The buyer alleged that the fibers shipped were from baby alpacas rather than llamas. Using microscopes, experts compared the diameter of fibers taken from the shipment to known samples of alpaca and llama. If the average diameters of alpaca and llama fibers were different, then the difference between the average diameter of the delivered fibers and the average diameter of known

12. 567 F.2d 1368 (8th Cir. 1977), *reh'g denied*, 567 F.2d 1381 (8th Cir. 1978), *cert. denied*, 439 U.S. 841 (1978).
13. *Id.* at 1375.
14. 330 F. Supp. 203 (S.D.N.Y. 1971), *aff'd*, 458 F.2d 1167 (2d Cir. 1972). *See also In re* Kroger Co., 98 F.T.C. 639 (1981).
15. 62 F.T.C. 1146 (1963).

fibers would have identified the shipment. This calculation of differences is only slightly more complicated than the discrimination cases examined above in that the average of observations made by the experts had to be calculated before the subtraction was made. The problem of deciding the significance of the difference between averages arises. If the average diameters of the two types of fibers are numerically close and fibers of a given type vary in diameter, can this difference still be used to identify the fibers?

This issue of the significance of the differences between averages arises more pointedly in *Presseisen v. Swarthmore College.*[16] A former assistant professor alleged sex discrimination in promotion practices by the college. The average time between receipt of the applicant's highest degree to promotion or appointment to assistant professor was 3.3 years for males and 5.8 years for females, a difference in averages of 2.5 years. Is this difference significant enough? The complication presented by *Presseisen* resembles that in the llama fibers case; this is not simply a summary of the difference between two absolute numbers such as expected and observed values. Rather it is a calculation of the difference between two averages, numbers that are already *summary statistics.* These statistics, like differences, abstract a particular characteristic from a larger group of numbers. It is not surprising that in determining the significance of differences between averages it becomes important to know the amount the individual numbers summarized by these averages vary from the averages themselves. If some men and women with identical characteristics are promoted in the same length of time, how significant is it that many others are not?

In other cases, litigators are concerned not with whether an observed number is significantly different from expected, but whether an observed number is significantly above a legal maximum or below a legal minimum. Often physical measurements combine the difficulty of comparing an average as in *Presseisen* with complexities introduced by a sampling process. For instance, in *Reserve Mining Company v. EPA,*[17] governmental agencies and environmental groups sought an injunction to prevent the defendant company from discharging wastes from its iron ore processing plant into the ambient air over Lake Superior. A court-appointed expert witness, Dr. William Taylor, testified that the concentration of asbestiform fibers was 0.0626 fibers per cubic centimeter of air, far below the legally

16. 442 F. Supp. 593 (E.D. Pa. 1977), *aff'd,* 582 F.2d 1275 (3d Cir. 1978).
17. 514 F.2d 492 (8th Cir. 1975), *modified,* 529 F.2d 181 (8th Cir. 1975).

permissible maximum for occupational exposure to asbestos.[18] To determine the difference between actual fiber concentrations and the legal maximum, Dr. Taylor measured the fiber concentration. Since airborne fiber concentrations vary on different days and in different locations depending on, among other things, the output of the iron processing plants, and meteorological conditions, Dr. Taylor averaged the fiber count of five testing sites. The mean (arithmetic average) from the five sample sites was compared to the legal standard.

The basic problem in *Reserve Mining* is the same as in *Castaneda*: determining the significance of a difference between two magnitudes. To evaluate the significance of the difference between sample means and other numbers, two factors must be taken into account: namely, the amount of variation between the individual items in the sample group and the sample mean, and also the size of the sample group.

This same process may be used to establish a legal maximum. In *Marathon Oil Co. v. EPA*,[19] one issue was the appropriateness of the manner by which the Environmental Protection Agency set its effluent discharge standards for the oil, mud, grease, and soaps that are washed from offshore oil drilling platforms. The EPA set these effluent standards by referring to the average performance of the best existing pollution control technology. The agency sampled discharges from some of the plants that were using that technology in an exemplary fashion. At issue in the case was whether this sampling method gave an emission standard significantly different from the level a plant using the best technology could reach 100% of the time. The disparity arose from complexities introduced into the measurement problem by the sampling and averaging processes.

In *United States v. General Motors Corp.*,[20] the Director of the National Highway Safety Bureau alleged that the automobile manufacturer failed to issue a safety defect notification to purchasers of an allegedly defective wheel used on light trucks. The relevant issue was whether the number of wheel failures was "significant" enough to require customer notification, essentially whether the number exceeded a minimum below which no notification is required. The difference between actual failures and the minimum could not be calculated directly because some customers fail to contact the company to report the defects. As in *Reserve Mining*, where the actual mean concentration of asbestiform particles was unknown, the number of wheel failures was estimated by sampling a selected group of customers. Since the number of wheels manufactured was known, the total number of failures could be estimated once the percentage of

18. *Id.* at 511 n.34.
19. 564 F.2d 1253 (9th Cir. 1977).
20. 377 F. Supp. 242 (D.D.C. 1974), *rev'd on other grounds*, 518 F.2d 420 (D.C. Cir. 1975).

wheel failures was determined. Calculating the difference between total failures and the minimum number triggering the notification requirement is straightforward.[21] Determining the significance of this difference is complicated by the sampling and averaging process.

There are simple differences between absolute magnitudes, between percentages, between magnitudes or percentages where there are numerous categories or groups involved, between averages calculated from measurements of all members of a group, between sample means, and between means and legal maximums and minimums. In each situation arising in a variety of substantive law areas, the basic problem is to determine the significance of a difference between values. The difference is basic subtraction. The significance of the difference is a statistical and legal question. The concept of a "difference" is the unifying method that binds all the statistical techniques involved.

Correlations

Subtraction is one measure of the relationship between numbers, specifically, the difference between them. Subtraction can be used only when the numbers are different measurements related to the same variable, such as expected and observed numbers of employees, or diameters of sampled fibers and known fibers, or actual discharge of effluent and legal maximum discharge of effluent. One could not, for instance, meaningfully subtract number of employees from total salaries, or diameter of llama fibers from sales price, or actual total discharge from the number of emitting facilities. There might, however, be a relationship between these pairs of variables that would be useful to summarize. It is likely that as the number of employees increases, the budget for salaries increases, or that as the diameter of llama fibers decreases, the sales price goes up, or that measured total effluent discharged will increase with the number of emitting plants.

In *United States v. City of Chicago*,[22] the court examined the relationship between the performance of police sergeants on a written promotion examination with their efficiency ratings on the job. Plaintiffs had proved that there were significant statistical disparities be-

21. Actually, the requirement of notification is triggered by evidence of a "large number" of failures. How large a "large number" is depends on the facts and surrounding circumstances of each case. Therefore, the legal minimum is the number below which the Director or, as in this context, the court, feels the number of failures is not large. 377 F. Supp. at 252 n.28.

22. 385 F. Supp. 543 (N.D. Ill. 1974).

tween Blacks and Whites selected to be promoted to sergeant, thereby shifting the burden to defendant to persuade the court that the written exam on which promotions were based was job-related.[23] One would expect that the sergeants with high efficiency and performance records would score well on a well-designed, job-related, written exam. If not, the score on the exam would be a poor predictor of future performance. When a high level of performance by an individual on one scale, e.g., actual performance, is matched with a high level by the same individual on another scale, e.g. the written exam, and an individual with low scores on one gets low scores on the other, it is said that the scores are *positively* or *directly correlated.* If sergeants with high scores on one scale had low scores on the other and vice versa, the scores would be *negatively correlated.*

The scores examined in this case showed a positive correlation for a sample of 176 incumbent sergeants.[24] Generally, though not always, those sergeants with high efficiency ratings scored higher on the written exams than those with low efficiency ratings. Once this conclusion is reached, it becomes important to decide how big or small the correlation must be to possess legal significance. In this case, how often will highly efficient sergeants be allowed to score poorly on the written test if one is still to conclude that the test is job-related? If 5% of sergeants have a negative correlation is that too many? 20%? 80%? The combination of statistical methods and substantive law explaining the significance of differences also addresses this issue; how often must a positive or negative correlation between the two measurements be observed in order to conclude that there is a significant positive or negative relationship overall? In *City of Chicago,* the court concluded that there was not a significant positive correlation overall.[25]

Determining the job-relatedness of a test can be complicated further by looking at correlations to actual performance, a process referred to as validation. In *Boston Chapter NAACP, Inc. v. Beecher,*[26] the Civil Service sought to validate use of its firefighters' exam by correlating exam scores with scores achieved by firefighters in each of thirteen job-related tasks such as ladder-extension, handling a pre-connected hose, air mask operation, extinguisher selection, securing lines and knots, and hose and hydrant operations. An expert witness, Dr. Hunt, testified that only two of the task scores showed a significant correlation to the exam scores and those corre-

23. *Id.* at 552.
24. *Id.* at 557-58.
25. *Id.* at 555.
26. 371 F. Supp. 507 (D. Mass. 1974), *aff'd,* 504 F.2d 1017 (1st Cir. 1974), *cert. denied,* 421 U.S. 910 (1975).

lations were "barely significant."[27] The court found that the Civil Service had not met its burden of demonstrating that the exam was "in fact substantially related to job performance."[28] The complication introduced by *Beecher* is similar to that arising in *Inmates* where dissimilar treatment of numerous groups was examined. From *City of Chicago* to *Beecher* a progression is made from examining a single correspondence, between scores and efficiency ratings, to numerous correspondences correlating fireman's written exam scores separately to their scores in each of thirteen job-related tasks. The greater the number of correlations that are calculated, the more often decisions as to the significance of the results arise. The intuitive judgment about the significance of the results becomes more difficult also.

The calculation of a correlation is admittedly more complicated than computing a difference, but the intuitive meaning of the concept is not difficult. The *correlation coefficient* is a summary statistic that describes the degree of correspondence between the values of two variables. The computation of the correlation coefficient always gives a value between -1, for perfect negative correlations, and 1, for perfect positive correlations. If the value of one variable, such as exam score, is totally unrelated to the measurement of the other, such as performance score, the coefficient equals zero. In *City of Chicago*, witnesses at one point reported a statistically significant correlation between police exams and performance of 0.247.[29] A statistician will say that there is only a 2% chance that this positive correlation would appear if there were really no relationship. This correlation of .247 is not very close to 1 which would show a perfect correspondence between examination and performance scores. In another case, *In re National Commission on Egg Nutrition*,[30] the Federal Trade Commission considered evidence of a correlation between average level of dietary serum cholesterol in each of seven countries and the corresponding number of new deaths from coronary heart disease and definite non-fatal heart attacks. This correlation coefficient was equal to .76 but was statistically significant only at a 5% level, indicating a 5% probability of random occurrence of this positive correspondence.[31] Why is .247 more significant than .76, which

27. *Id.* at 517.
28. *Id.*
29. 385 F. Supp. at 558.
30. 88 F.T.C. 89 (1976).
31. *Id.* at 130.

appears to be closer to a perfect positive correlation?

Just as a greater difference between two numbers should show a greater and, hence, more significant disparity, one would expect a correlation coefficient closer to one to be interpreted as showing a closer and, hence, more significant correspondence between the values of the variables. As will be seen, however, one key is the number of observations made of the variables in question. Not surprisingly, within certain limits, the more measurements made of a certain event, whether the event is selection of jurors, concentration of air pollutants, rates of coronary heart disease, or correspondences between test scores, the more reliable the conclusions will be, and the more confident the inference that there is a relationship between the variables will be. The number of observations will not necessarily increase the correspondence, but it will make a given correlation more reliable. Hence, a more significant input into our legal fact-finding process will exist. In *City of Chicago*, the correlation was calculated using 176 test scores. In *Egg Nutrition*, only seven countries were observed. This large difference in sample size affects the reliability of the results.

Knowing the degree of correlation between measurements of different variables is useful to explain how the value of one variable changes when there is a change in the value of the other. The greater the correlation between variables, the more the changes in one correspond to changes in the other. In *Egg Nutrition*, the increased risk of heart attacks and heart disease in men was examined by exploring its correlation with their consumption of eggs. The trade association of egg manufacturers would have loved to have been able to show a zero or even negative correlation between heart disease and the dietary intake of serum cholesterol, which comes from eggs. An expert witness, Dr. Connor, reported a correlation coefficient of 0.666 between the coronary heart disease death rate in thirty countries and the associated average daily intake of eggs.[32] To find out how much of the variation in heart disease among countries is mathematically explained by variations in egg consumption, one simply squares the correlation coefficient. Thus 44% (0.666 x 0.666) of the variability in heart disease rates is accounted for by variability in egg consumption, according to Dr. Connor.[33] Other testimony revealed that 66% of the variation in the level of new diagnoses of coronary heart disease among men in seven countries and 58% of variation in deaths from coronary heart disease and definite non-fatal heart attacks were accounted for by the variation in mean level of dietary serum cholesterol; correlation coefficients of 0.81 and 0.76

32. *Id.* at 133.
33. *Id.*

respectively.[34] How much of the variation must be accounted for by eggs in order that the trade association's advertising that eggs are good for you becomes deceptive and misleading? Is 44% enough? 58%? 66%? The significance of the amount of variation in one variable accounted for by variations in the other is another example of the conjunction of statistical methods and law.

It is noteworthy that in order to use correlation coefficients one need not have numerical measurements, a so-called *cardinal ranking* which tells by how much one measurement is bigger than another. It is enough to be able to place the different observations of variables in a list ranking the observations in order from top to bottom, an *ordinal ranking.* In *Commonwealth of Pennsylvania v. Local Union 542, International Union of Operating Engineers,*[35] the court reports such an approach. A class of minority workers alleged employment discrimination in the operation of the union's referral system. The statistical expert prepared two ordinal lists. The first ranked members of the union by how often they were out of work while the second ranked members in order in which they were actually referred to jobs. The theory was that members out of work longest would be referred to jobs first if the assignments were made on a non-discriminatory basis. The rank orderings of workers on referral and out-of-work lists were compared by calculating a correlation coefficient. A perfect positive correspondence would give a correlation coefficient of 1 and would be evidence of a non-discriminatory referral system.

The expert prepared seventeen pairs of lists representing different job categories and time periods.[36] Correlation coefficients ranged

34. *Id.* at 130.

35. 469 F. Supp. 329 (E.D. Pa. 1978), *aff'd,* 648 F.2d 922 (3d Cir. 1981), *rev'd sub nom.* Gen. Bldg. Contractors Ass'n, Inc., v. Pennsylvania, 102 S. Ct. 3141 (1982).

36. The resulting correlation coefficients and percentage of variance explained by rankings for the seventeen lists were as follows:

List Number	Rank Correlation Coefficient	Percentage Variance Explained	Percentage Variance Not Explained
1	.24	5.8	94.2
2	.08	0.6	99.4
3	.22	4.8	95.2
4	.20	4.0	96.0
5	.40	16.0	84.0
6	.38	14.4	85.6
7	.55	30.3	69.7
8	.44	19.4	80.6
9	.37	13.7	86.3
10	.52	27.0	73.0
11	.44	19.4	80.6

from .08, showing almost no relationship between the lists for a given time period and job category, to 0.62. The out-of-work lists, which, if used as a guideline for referral, would explain the orderings on the referral lists from 0.6% to 38.4% of the referrals. This left as much as 99.4% and as little as 61.6% unexplained. The court found that this corroborated the plaintiffs' claims of discrimination in that it proved there was much room for arbitrary and standardless selections: "When combined with the other statistical disparities considering the race factor directly, this correlation study aids the inference of discrimination."[37] The statistical expert concluded that virtually none of the lists reflecting actual referral rankings were *significantly similar* to the corresponding out-of-work list. As will be seen, the significance of a correspondence is determined in a fashion similar to the significance of a difference.

Regression and Multiple Correlations

Regression

Correlation coefficients summarize the extent to which the measurements of two variables change in the same direction at the same time. They do not, however, allow one to predict *how much* the measurement of one variable changes when there is a change in the value of another variable. Related to the correlation coefficient is the *regression coefficient*, which does provide this information. In *South Dakota Public Utilities Commission v. Federal Energy Regulatory Commission*,[38] the state utility commission opposed a FERC order permitting an accelerated rate of depreciation for certain facilities owned by Northern Natural Gas. The FERC order was based on a finding that because reserves of natural gas were dwindling, Northern's equipment would lose its value before its physical life was over. One study designed to project future reserves was based on a theory that related drilling efforts to results. The relationship between two variables, time and new discoveries, was examined to project a year in which no new discoveries would be made. The regression coefficient was -160.16 billion cubic feet per year.[39] This meant that the first variable, new annual discovery, decreased by 160.16 billion cu-

12	.43	18.5	81.5
13	.46	21.2	78.8
14	.46	21.2	78.8
15	.54	29.2	70.8
16	.62	38.4	61.6
17	.45	20.3	79.7

Id. at 356.

37. *Id.* at 357.

38. 643 F.2d 504 (8th Cir. 1981), *rev'd on other grounds*, 668 F.2d 333 (8th Cir. 1981).

39. *Id.* at 511 n.10 (this figure is calculated from data in the case).

bic feet of gas for a one year change in the second variable, time. Given the current level of discoveries, calculations projected that no new discoveries would be made after 1981.

The reliability, accuracy, and, hence, significance of this projection will depend, not surprisingly, on the basis from which the prediction is derived. One would not predict winter snowfall for New England by reference only to Arizona's weather, or by looking at just one or two years of New England's weather history. Nor would one attach much significance to a precise estimate of this year's snowfall if historically a wide variation in snowfall from year to year were observed. The logical relationship between the variables, the number of measurements made, and the variability among those measurements will all affect the significance of the conclusion. Statistical tests of each of these parameters guide the fact-finder in assessing the significance of the conclusion.

For regression and correlation coefficients it is important to know whether the results are due to some fluke or aberration in the numbers used to calculate them and whether the size of the coefficient is large enough to be legally significant. Both coefficients are calculated in order to summarize the correspondence between two variables. If there is no numerical correspondence, the values of each coefficient will be zero, though the regression coefficient is not constrained to fall between -1 and 1 as the *South Dakota* example illustrates. From a statistical viewpoint, the question of significance is whether the calculated coefficient is truly different from zero or, to put it another way, whether any difference is due merely to some fluke or aberration or chance. This statistical issue of significance is conceptually identical to determining whether the *difference* between expected and observed numbers is large. For the discrimination, contract, and environmental cases described in the previous section, we continually questioned the significance of a difference—between expected and observed, between observed and a legal maximum or minimum, between averages from different samples. Is a difference large or is it truly no different from zero, the result of some random chance? Calculating percentages and averages may be mathematically easier than calculating coefficients, but the process of determining their statistical significance is roughly identical. And the legal significance of a particular fact is a practical question most familiar to the litigator. The legal significance is a question of materiality and relevance; the statistical significance is a question of persuasiveness, or weightiness of the evidence.

The particular utility of regression coefficients is that one is able to estimate simultaneously the individual effects of numerous explanatory variables on the variable one is trying to estimate, predict, or explain. The variable sought to be explained is referred to as the *dependent variable* because its value depends on the values of the explanatory variables. The explanatory variables are referred to as *independent variables*. In *Presseisen v. Swarthmore College*,[40] one expert tried to predict salaries of college teachers (the dependent variable) by summarizing the effects of sex, age, years since highest degree, years teaching at Swarthmore, degree, and academic division (the independent variables). An opposing expert testified that other variables such as scholarship, teaching ability, publications and their quality, quality of degree, duration of career interruptions, and administrative responsibilities, also would affect salary and should have been included in the regression calculations. In a typical race discrimination in employment case the plaintiff's regression analysis might include the following independent variables to explain salary differentials: minority status, years of education, years since receipt of highest degree, age of employee, years of prior experience, and years of employment by the defendant employer.[41]

No matter how many variables are simultaneously calculated, the coefficients can be examined one by one. By one expert's testimony in *Presseisen*, for instance, the coefficient describing the relationship between salary and sex in 1972 had a value of -340, indicating that women teachers on average made $340 less than men with comparable skills.[42] This is the coefficient of interest in a sex discrimination suit. A familiar problem arises, however, in determining the practical and statistical significance of this difference between male and female salaries. Is $340 significantly different from zero given the absolute size of salaries? Not surprisingly, the significance will depend on the number of measurements taken and the variation among those measurements. If men's salaries differ widely from one another it would not be as surprising to find women's salaries differing on average by some small amount from men's.

Nor is the use of multiple regression coefficients limited to discrimination cases. The ability to explain variations by districts in per pupil expenditures on education was critical to defenders of the Washington state school financing system examined in *Northshore School District No. 417 v. Kinnear*.[43] Mr. Francis Flerchinger, a

40. 442 F. Supp. 593 (E.D. Pa. 1977).
41. Agarwal v. Arthur G. McKee and Co., 19 F.E.P. Cases 503 (N.D. Ca. 1977).
42. 442 F. Supp. at 617.
43. 84 Wash. 2d 685, 530 P.2d 178 (1974), *rev'd on other grounds sub nom.* Seattle School Dist. No. 1 of King County v. State of Washington, 90 Wash. 2d 476, 585 P.2d 71 (1978).

statistician from the Office of Superintendant of Public Instruction, tried to explain the level of expenditures per pupil in each of Washington's 320 school districts by reference to two variables: teachers' pay and certified staff per thousand students in each district. The coefficient on each of these two independent variables describes the effect of variations in the variable on the level of expenditures per pupil just as sex or race might explain salary differences among teachers. Again, the reliability of the coefficients would have to be tested.

The regression coefficient is conceptually a combination of the difference between averages and the correlation coefficient. The coefficient of -340 associated with the sex variable in *Presseisen* is the average difference between male and female salaries. Because the correspondence between sex and salaries is being examined, however, the regression coefficient explains how the value of one variable changes when there is a change in the value of the other, just as a correlation coefficient does. A correlation coefficient summarizes the correspondence between values of two variables; *whether* one variable increases in value as the other does: whether, as ingestion of dietary serum cholesterol increases, the rate of coronary heart disease increases. A regression coefficient adds information as to *how much* the value of one variable increases or decreases when the other changes. The coefficient of -340 tells not only that as sex changes from male to female, salaries decline, but also that the decline is $340. In *South Dakota Public Utility Commission*, the coefficient of -160.16 tells not only that new discoveries are declining. It also tells that as each year passes, that is, the variable time increases by one year, the new annual discoveries of natural gas decrease by 160.16 billion cubic feet from the level of the year before.[44]

Multiple Correlations

Multiple regression is often used to predict or explain the value of a dependent variable such as new discoveries, salary, or expenditures per pupil. There are numerous independent variables such as sex, experience, age, or certified staff per thousand students. Therefore, it would be helpful to know how accurate a prediction or complete an explanation is once the separate influences of all the various relevant explanatory variables have been taken into account. Squaring the correlation coefficient in *Egg Nutrition* indicated how much varia-

44. 643 F.2d at 511 n.10 (this figure is calculated from data in the case).

tion in heart disease among various countries is accounted for by egg consumption. This *squared* coefficient varied in value between 0, for none of the variation explained, and 1, for all of the variation explained. The *coefficient of multiple determination* in *Northshore School District* was reported by the court to be .75, indicating average pay for certified staff and staffing ratio per 1,000 pupils combined to account for 75 percent of the variation in expenditure per pupil. The coefficient of multiple determination takes the explanatory power of all the variables into account. As with the simple correlation coefficient, it is important to determine whether the amount explained produces trustworthy estimates.

EVALUATING THE SIGNIFICANCE OF NUMERICAL CALCULATIONS

Standard Deviations

For many of these cases and for most relevant statistical tests involving large samples, the key notion is the *standard deviation.* Like the average, the standard deviation is a summary statistic.[45] It measures how much a typical observation varies from the average of all observations. It is useful in probability because it is more likely that one will be confronted by an unusual observation if there are many unusual events than if there are few. Law and statistics are concerned with the probability of unusual events occurring just by chance because of an interest in causation.

The Significance of Simple Differences

The criminal defendant's claim in *Castaneda*[46] was that the jury selection process was biased against the selection of Hispanics and this resulted in an underrepresentation of that group. Bias is not the only logical explanation, however. The underrepresentation could also have occurred by chance. For instance, it is unlikely that all the judges in this county were born under the same astrological sign, but that phenomenon could occur by chance. It is unlikely that a fair coin could be flipped or a juror could be randomly selected one thousand times, resulting in a head or a Hispanic only fifty times, in a county where Hispanics are a majority of the population; but it could occur by chance. If the frequency of such unlikely events occuring could be calculated, then the unlikelihood of a particular event could be appreciated. The "event" is the absolute difference

45. For some statistical tests, particularly those involving small numbers of observations, mathematical formulas are used to calculate the probabalistic information revealed by the standard deviation. *See* D. BARNES, J. CONLEY, AND D. PETERSON, *supra* note 2 at Section 4.6 *et seq.* The same interpretations are attached to those results, however, so the following discussion applies to them as well.

46. 430 U.S. 482 (1977).

between the observed number and the expected number, also known as the norm. For example, given half heads/half tails for a fair coin or 79.1% Hispanics in the eligible population, the norm would be 79.1% Hispanics on the jury. How likely is it that one would actually observe only 39% Hispanics just by chance, that is, without any bias in the selection procedure? This is a measure of the probability that it would be wrong to conclude that there was bias in the system.[47] A high probability of an observed disparity occurring by chance is an alternative and exculpatory explanation of causation.

What factors affect the probability of a random occurrence? For all the measures of differences and correspondences, two primary factors are relevant, though for different statistics they combine in different ways.[48] The first factor is the size of the group about which or from which one is attempting to draw particular probabilistic conclusions. For many jury discrimination or employment discrimination cases where absolute or percentage differences are the relevant evidentiary inputs, this group is the *group of suspect composition*, the number of jury members discriminatorily selected by suspicious means, the workforce that was hired by someone who allegedly took improper criteria—race, sex, alienage—into account. For cases such as *Presseisen*, comparing the difference between average time to promotion or average salaries for two groups, the sizes of those groups are relevant. Where only part of the universe of measurements are observed, such as in *Reserve Mining*, and for correlation coefficients and regression coefficients, where correspondences between values of numerous variables are calculated, the number of observations of relevant variables actually made and, in the case of regression, the total number of variables are relevant. Why?

Consider the probability of a particular occurrence happening by chance. If two people are randomly selected from their downtown offices during business hours and locked in a room, it would not be surprising if neither was a lawyer. If twenty people were randomly selected it would still not be surprising if no lawyer was among them even given the supposed glut of attorneys. If three hundred offices were randomly raided, however, certainly we would casually and unscientifically agree that one was likely to be a law office and that we

47. In statistics the probability of finding an innocent person guilty is called a type one error. A type two error involves finding a guilty person innocent.

48. Another important factor is the distribution of observations within the population from which a sample is drawn. Probabilities for these samples are dealt with by mathematical formulas if the samples are small and by standard deviation approaches if they are large.

could lock up a lawyer. So the size of the sample or of the group of suspect composition is relevant to the determination of the probability of random occurrence. And note that it does not matter how big a city is used for the kidnapping. The absolute number of law offices is not relevant, only the percentage of total offices that hide lawyers. This logic applies in a straightforward manner across all statistical tests.

The second common factor is the diversity of objects within the class or classes being examined. The number of minority members on a jury or a workforce will depend not only on the size of the group selected but also on the racial composition of the population from which the group of suspect composition was selected. One expects a higher proportion of minorities to be selected from a population that has more minorities, regardless of the total size of that group. The larger a minority group is, in percentage terms, the more diverse is the population as a whole.

The concept of the standard deviation combines a statistical measure of diversity with a measure of group size. For a population with greater diversity, the standard deviation, which may be thought of as the typical or average deviation from the norm, mean, or expected value, gets larger. If the largest group in the population is white, the standard deviation is larger as the percentage of blacks, reds, and yellows gets larger. If a large proportion of llama fibers measured has a diameter quite different from the average diameter, then the standard deviation (or "typical" deviation of llama fiber diameters from the average) will be greater than if all of the diameters are quite similar. In the llama fiber case, in fact, expert witnesses compared the standard deviation of llama fibers from their average diameter to the standard deviation for alpaca fibers.[49] Because llama fibers vary more from one another in diameter (there is more diversity) the standard deviation calculation enabled the scientists to identify the fiber stocks of the defendant seller as llama fibers even though they had a similar appearance and identical average diameter to alpaca fibers.

The Significance of Differences Between Averages

The standard deviation is also relevant when comparing averages of several groups, such as the length of time to promotion for men and women in *Presseisen.* The standard deviation for each group is an indicator of the reliability of an estimate of the absolute difference between the averages for men and women. If one were to predict that a typical male would be promoted in 3.3 years, give or take a year, and that a typical female would be promoted in 5.8 years,

49. 62 F.T.C. 1146, 1148 (1963).

give or take two years, one would not only identify a greater diversity among women ("give or take two years" — indicating less precision about the prediction), but note that some women who are promoted quickly are promoted before some men who are promoted late. The amount one must "give or take" in the prediction depends on the diversity or variation within each group and may, if it is large, indicate that there really is no significant difference between the length of time for promotion of men and women after all.

The Significance of Correlation and Regression Coefficients

Standard deviations are the key to significance testing not only for the calculations of differences but also the measures of correspondence between two or more variables, the correlation and regression coefficients. Because the measures of correspondence involve grouped comparisons of values of two or more variables, the diversity of values for each variable is taken into account in determining the statistical significance of the numerical result. The correlation coefficient can be calculated, in fact, by comparing whether and by how much the paired values of the two variables simultaneously differ from their respective means. Thus, in *City of Chicago*, comparing written exam performance to job performance, the correlation coefficient involved a determination of whether each police sergeant who scored well above average on the written test also scored well above average on the performance rating. If there is a great amount of variation in the correspondences, the correlation coefficient is less likely to be statistically significant.

Because regression coefficients are used to predict the quantitative influence of one variable on another, it is useful to compare the observed values of the variable to be explained to the predictions given by the regression coefficients. The more the observed and predicted values deviate from one another, the less reliable the estimate.[50] Thus, if the typical deviation of actual from predicted value is 1000 billion cubic feet (BCF) for the predicted new annual discovery of natural gas in *South Dakota Public Utilities Commission*, and the average new discovery is 1531 BCF, one would not want to rely on

50. In statistical terms this standard deviation is referred to as the standard error of the estimate. In fact, as a matter of terminology, whenever the standard deviation is computed from a sample rather than from the entire population, it is referred to as a standard error rather than a standard deviation. It might be the standard error of the mean, or the standard error of the estimate, or the standard error of the sample, depending on what is being calculated. But whichever standard error is involved, the underlying goal is to find some measure of the typical deviation from some norm or mean value.

the estimate because the uncertainty is so great.

Why The Standard Deviation Is Useful

An attempt is made to answer the questions "How big a difference or correspondence is big enough?" and "Under what circumstances is big significant?" Big will be significant if a number of that size is unlikely to have occurred by chance, that is, has statistical significance, and if a number of that size has legal implications, that is, has practical significance. The lawyer supplies the expertise as to practical significance while the statistician supplies the expertise with respect to statistical significance. The purpose is to establish or eliminate chance as an explanation for the phenomenon giving rise to the legal dispute. The standard deviation, by indicating a typical variation from the norm, provides a yardstick for measuring how atypical a particular observation is. An atypical observation may have an exculpatory explanation such as chance — the defendant just happened to be walking by the bank as it was robbed or just happened to hire no minorities — or an inculpatory explanation — the defendant intended to rob the bank or intended to hire no minorities. How credible is the bad luck explanation?

To address the question of how "atypical" is "atypical", it is important to decide what is typical. The definition will involve the probability that a particular observation will occur by chance. If that probability is high then the observation is a typical one. If the probability of an observation occurring by chance is small then the observation is atypical. For many magnitudes to be measured, it is a useful mathematical fact that approximately two-thirds of all random observations will be within one standard deviation of the mean or expected value. In *Castaneda*, the court calculated the standard deviation as being equal to 12 jurors.[51] Given that the expected number of Spanish-surnamed jurors in a given time period was 79.1% of the 870 jurors or 688 jurors, it is expected that if juries are drawn randomly, that is, without discrimination, that two-thirds of the time there would be 688 plus or minus 12 Spanish-surnamed jurors, that is, from 676 to 700. Approximately 95% of juries actually chosen randomly will be within two standard deviations of 688, that is, 688 plus or minus 2 times 12, and 99% of juries will be within three standard deviations. The actual jury history in *Castaneda* showed 339 Spanish-surnamed jurors, an observation that was more than 29 standard deviations from the expected number of 688 (688 - 29 times 12 = 340). An observation more than one standard deviation from the mean occurs about once in every three times (about 33% of the time) because observations less than one standard deviation away

51. 430 U.S. at 496 n.17.

from the mean occur two-thirds of the time. An observation more than two standard deviations away occurs only one time in twenty while an observation three standard deviations away occurs by chance only one time in a hundred. An observation 29 standard deviations away occurs only once in every 10^{140} times; that is a ten with one hundred and forty zeros after it. Now, that is a big number and a very small probability of chance occurrence.

The Supreme Court in *Castaneda* gave tacit recognition to the statisticians' standard that an observation occurring less often than once in twenty or once in a hundred times is atypical and unlikely to be due to chance alone.[52] There is no particular reason why this probability should be chosen and it may be inappropriate for law where serious consequences attend a conclusion based on probability. Nevertheless, courts often say that a difference between expected and observed values of more than two or three standard deviations is statistically significant. For these observations, it can be said that there is only a 5% or 1% chance, respectively, that the observation occurred by chance alone.

Tests based on the statistical properties of numbers that give rise to these characteristics of the standard deviation can be applied to the variety of statistical problems described above. For the various summary statistics calculated, statistical tables or formulas show the probability that such a number could occur by chance. If that chance is less than one in twenty, statisticians often conclude that chance is not a significant causal explanation. It is worth repeating that the one in twenty threshhold (or 5% significance level) is merely a convention adopted by some physical and social scientists. While some courts have adopted it, they have regularly done so without examining the propriety of adopting this level in an application such as law where serious consequences attach to findings of fact.

Significance Testing

Using statistical tables or formulas, conclusions can be drawn about the significance of the numerical results presented in the previous cases. For each of the cases discussed, a statistical test indicates how big is big, how big a difference or correspondence has to be to be believable and significant, how reliable a prediction is, or how significant is a particular explanatory variable. For each statistical tool, the outcome depends on the variation of observations around

52. 430 U.S. at 494-96.

the mean value, that is, some version of the standard deviation, and some measure of the size of the sample or group of suspect composition. The statistician, after calculating the test statistic, checks a table or employs a formula which shows, for varying sample or group sizes or number of categories being examined, the likelihood that the difference, correspondence, or prediction occurred by chance and, hence, is unreliable or statistically insignificant or significant.

Simple Differences

In the *Castaneda* and *Goff* cases, the respective differences in percents of 40.1 percentage points and 5.27 percentage points were easily distinguished. Such a large disparity as 40.1 percentage points is unlikely to occur by chance while the 5.27 percentage points is comparatively rather small. Why then, in *Board of Education*, is 4.2 percentage points statistically significant?

An appropriate test statistic for determining the statistical significance of some simple differences between two numbers or percents is the Z statistic, calculated by dividing the difference between the numbers or percents, i.e., the disparity, by the standard deviation. The Z table indicates how probable it is that the disparity could occur by chance, i.e., whether it is a statistically significant difference. In *Board of Education*,[53] the disparity of 4.2 percentage points gives a Z test statistic of -2.52. The Z table reveals an associated significance level of .0118 indicating that in this group of teachers, the observed underrepresentation of Blacks is likely to occur by chance only 1.18% of the time if the hiring process is really done without reference to race.

For other differences such as the disparity in *Goff*,[54] a mathematical equation called the binomial formula is used to calculate the significance level directly. The formula is used because the group size under consideration in *Goff* was rather small, a jury of 23 members, compared to the group size in *Board of Education*, a school with 129 teachers. There are also tables that help avoid long calculations. The binomial formula indicates a significance level of about .42, indicating that the probability of finding a disparity as large as the one that occurred here just by chance was about 42%. Compared to the 1.18% probability in *Board of Education*, the 42% probability makes chance rather than discrimination appear to be a more likely explanation for the disparity.

Thus, the disparity in *Board of Education* may be found to be statistically significant, that is, not likely to have occurred by chance, while the disparity in *Goff* based on a group size of 23 is not statisti-

53. 584 F.2d at 584-85 n.29.
54. 509 F.2d at 827.

cally significant. By now the mystery of statistics may have been cleared up as to the question of how to calculate the test statistics and how the tables work, but for those explanations the reader is referred to texts and treatises on the subject.[55]

Differences with Numerous Categories

Disparities involving two or more categories can be analyzed using a similar procedure. Recall in *Inmates*, for instance, there were disparities in four different racial groups being examined simultaneously. An appropriate test statistic for this case is the *chi-square* statistic.[56] The chi-square table shows, for various numbers of categories, the probability of getting different values of the chi-square statistic by chance. For these cases the probability associated with the chi-square test statistic indicates the probability of different treatment among categories.

In *Certified Color Manufacturers Ass'n v. Mathews*,[57] the chi-square test statistic for two categories, high dosage and low dosage rats, equals 1.85.[58] The chi-square table indicates that, given this chi-square statistic, there is between a 10% and 20% chance that one might observe the same difference in cancer rates just by chance even if the high dosage of Red Dye No. 2 has no more of a carcinogenic effect on rats than the low dosage. Here, the significance level of .10 to .20 corresponds to a 10% to 20% chance of random occurrence. These values of .10 and .20 are also referred to as *p*-values. Because the standard referred to in this case required a significance level of 5% or lower and *p*-values of .20 to .10 correspond to between 20% and 10% significance levels, this finding was not thought statistically significant.

In *Inmates*, a chi-square test for the four categories gives a value of 7.49[59] indicating between 5% and 10% probability that the allegedly discriminatory result of the discretionary parole hearings was due to chance, a 5% to 10% significance level. Because the signifi-

55. *See supra* notes 1 and 2.
56. Chi is pronounced Ki to rhyme with sky.
57. 543 F.2d 284 (D.C. Cir. 1976).
58. This figure is calculated for an observed 7 of 23 high dosage rats and 4 of 30 low dosage rats developing cancer to determine whether there is a significant departure from the mean cancer rate of 21%. *See* 543 F.2d at 290 nn.30-31.
59. This figure is calculated for observed numbers of 358 of 590 Whites, 184 of 235 Blacks, 24 of 59 Native Americans, and 5 of 18 Mexican-Americans receiving discretionary parole to determine whether there is a significant departure by racial group from the mean discretionary parole rate of 59.3%. 567 F.2d at 1371.

cance level was greater than 5%, among other reasons, the court concluded that there was no proof of discrimination sufficient to make out an equal protection claim.[60]

In *Chance*, the court found that the probability that the small actual number of minorities passing the exam was due to random factors alone was less than one in a million,[61] clearly a statistically significant result by conventional standards.

When disparities are statistically significant, the statistician concludes that the difference is big. When a difference is not statistically significant, when there is a sizable probability that it could occur by chance, it is concluded that the difference is small, whatever its absolute size may be.

Differences in Samples

When numerical calculations result from samples rather than from a measurement of the whole population as in *Reserve Mining*, *Marathon Oil*, and *General Motors*, an uncertainty is introduced. An estimate or prediction is less credible than actual knowledge. To adjust for this uncertainty, statisticians use a variation of the standard deviation called the *standard error of the mean* and the *t* table, described above, to indicate a range around the estimate where it can be reasonably sure the true value lies. This range is called a *confidence interval.* Thus in *Reserve Mining*, the scientist estimated an airborne asbestiform fiber concentration from his samples of 0.0626 fibers per cubic centimeter (cc) of air with a 95% confidence interval of plus or minus 0.0276 fibers per cc.[62] This meant that there was a 5% chance that the actual concentration lay outside the range of 0.035 fibers per cc (the low end of the interval) and 0.0902 fibers per cc (the high end of the interval). In *Marathon Oil*, where the EPA sought to state a legal maximum of allowable pollution from offshore oil drilling rigs by estimating how much effluent would wash off sample high technology rigs, they took the top of the confidence interval which they calculated around the estimated average level of effluent in their sample. The top of the interval was used to set the legal maximum because then the EPA could be certain that average emissions from the most modern facilities would be above this level less than 1% of the time. In *General Motors*, by contrast, the court sought to determine whether the number of wheel failures was greater than the legal minimum necessary to trigger the defect notification procedures.[63] In this case, the low end of the interval

60. *Id.* at 1381.
61. 330 F. Supp. at 210.
62. 514 F.2d at 511 n.34.
63. 377 F. Supp. at 245.

around the estimated number of reported failures was used because the government wanted to know whether the actual number of failures was above the legal minimum.

Differences Between Averages

In *Presseisen*, where average time in rank for men and women was computed to determine whether women are discriminated against in academic promotions, one test of the significance of the difference between these averages is to determine whether the confidence intervals associated with each average overlap. If they do, then there is significant probability that the average times in rank for the two sexes are not really different.[64] Not surprisingly, plaintiff's expert in *Presseisen* stated that he was unable to test the statistical significance of the differences between time in rank.[65] Computations of confidence intervals assuming hypothetical yet reasonable values for the numbers of individuals in each rank show that for every rank from instructor through full professor there is an overlap of the intervals around the separate estimates for men and women. The court appropriately found no proof of discrimination in promotion.[66]

Correlation Coefficients

The correlation coefficient, given the statistical symbol r, is the test statistic for the measure of correspondence. One refers to an r table which shows, for different sample sizes, the statistical significance of the correlation coefficient. If an r value is statistically significant, then there is a relatively small probability that the correspondence observed between two variables occurred by chance. It should come as no surprise by now that one correlation coefficient may be larger than another in absolute magnitude yet less significant statistically because it is based on a smaller sample. A statistically significant correlation coefficient, r equal to 0.247, was noted in *City of Chicago*, and another, euql to 0.76, in *Egg Nutrition*, was also statistically significant. The *City of Chicago* study was based on the test scores of 176 police sergeants and the r reveals an associated

64. 442 F. Supp. at 612.

65. There are a variety of statistical tests for the significance of differences between sample means many of which are based on the t test, as is the confidence interval approach. In *Presseisen*, the averages taken from all actual faculty promotions are tested as an estimate under the theory that the history of actual promotions of Swarthmore is only a sampling of the administration's behavior over the long run. *See* 442 F. Supp. at 609.

66. *Id.* at 612.

significance level of 2%. So it can be said that there is just a 2% chance that positive correlation between written exams and performance scores of this size could occur just by chance, that is, just because of the particular items observed. In the *Egg Nutrition* example, with a sample size of seven countries, the significance level associated with the calculated correlation coefficient of 0.76 is 5%. In this case it can only be said that there is a 5% chance of random occurrence of this correspondence.

The same logic applies to the rank correlation coefficient although a different table must be used. Sample size will influence the statistical significance of a particular coefficient. Thus for *Local Union 542*, if the out-of-work list and referral list being compared have 30 union members listed on them, a calculated r of .45 will be statistically significant at the 2% significance level while if they have only 6 union members on them a calculated r of .45 would not be statistically significant even at the 10% significance level.

It does seem peculiar that a correlation of .247 from one set of data would be more significant than a correlation of .76 from another set, but remember that this is only the *statistical* significance of this result. The litigator's judgment is appropriate to determine the persuasive value of the statistical evidence. It depends on the other proof available and the nature of the factual issues involved in a particular lawsuit. The process of squaring the correlation coefficient, r, to get the percentage of variation in one variable explained or accounted for by correspondence with the other variable, called the *coefficient of determination*, r^2, discussed in the *Egg Nutrition* case, is a guide to practical significance. Even though an r of 0.249 in *City of Chicago* is statistically significant, only 6.2% of the variation in written scores was accounted for by the correspondence to performance scores. The use of the correlation coefficient is best thought of as a two-stage process: first, determining the statistical significance by reference to the r table, and second, determining the practical significance by reference to the litigator's knowledge of the factual issues involved and to the coefficient of determination, r^2, to see how much explanatory power the variables really have.

Regression Coefficients

Significance testing for multiple regression does not involve any complex new techniques. The practical utility of regression results depends on the reliability of predictions produced. The reliability of the predictions depends on two related factors. The first is the cumulative explanatory power of the independent variables. The second is the accuracy of the estimates of each independent variable's influence on the dependent variable. In *South Dakota Public Utilities Commission*, then, one wants to know whether the regression coeffi-

cient of -160.16 billion cubic feet is big enough to be statistically different from zero. If it is not then one cannot be reasonably certain that there is any predictable change in additions to annual reserves of natural gas (the dependent variable) over time (the explanatory or independent variable). The test procedure is similar to that for testing whether a disparity between numbers or percentages is big, as was done in *Castaneda*, *Goff*, and *Inmates*. Here the disparity is between the calculated value of the regression coefficient, -160.16 billion in *South Dakota Public Utilities Commission*, and zero. We are testing whether this coefficient is significantly different from zero just as we tested whether the observed number of a minority group was different from the expected number. To use the *t* test in either case, divide the disparity by the standard deviation. For regression coefficients, the equivalent to the standard deviation is called the standard error of the regression coefficient. The division in this case gives -2.209, which one compares to values on the *t* table for the correct sample size, ten years in this case. The *t* table reveals that the division must yield a number whose value, ignoring any minus sign, is greater than 1.860 to establish that there is less than a 5% probability that this estimate of the annual decline in results from gas exploration is less than zero only by chance. Since the calculated value is greater than 1.860, the critical value, the regression coefficient is statistically significant by this 5% standard and we would be willing to accept this statistical evidence of a yearly decline in annual additions to gas reserves.

When, as in *Presseisen*, more than one variable is used to explain a dependent variable, the *t* test applied separately to the regression coefficient for each independent variable separately indicates the statistical significance of each variable's coefficient. Thus, a *t* test applied to the regression coefficient of -340 dollars on the variable "sex", one of a number of variables used to explain variations in faculty salaries, would only allow us to draw inferences about whether sex had a significant part in the salary determination process. It would not indicate whether all of the variables considered, taken together, give a reliable salary prediction. In *Presseisen*, Dr. Iverson, an expert witness for the defendant, testified that his estimate of salaries as $340 lower for women than for comparable men was not statistically significant.[67] Dr. Iverson testified that the probability that the negative values for the coefficient on the sex va-

67. *Id.* at 613.

riable would occur by chance varied from 37% to 66% for different years in which the coefficient was calculated. For no year did the probability of random occurrence drop below the 5% significance level.[68] The court refused to find that sex had a discriminatory impact on salary determination.[69]

When the emphasis shifts from the statistical significance of each independent variable to the reliability of the prediction from all the variables taken together, an F test is substituted for the t test to determine the statistical significance of the calculated coefficient of multiple determination, R^2, which, you will recall, indicates the percentage of variation in the dependent variable that is accounted for by variation in all the independent variables. Like the other test statistics, calculating the F statistic includes an equivalent of the standard deviation called the *standard error of the estimate* which determines the typical deviation of the predicted values from the actual observed values of the dependent variable. In addition, it measures the standard deviation of the observed values of the dependent variable from their average value. The calculated F statistic is then compared to an F table which, for varying sample sizes and numbers of independent variables, shows the reliability of the prediction in terms of the cumulative statistical significance of all the explanatory variables. The significance level associated with a particular calculated F indicates the probability that we would observe as great an explanatory power as is indicated by the coefficient of multiple determination if there were no actual relationship between the dependent variable and any of the independent variables. Thus, in *Northshore School District* a calculated F value of 475.7 was highly significant because the F table indicates that the critical value for a 5% significance level and the sample size of 320 is about 3.86. The calculated F of 475.7 exceeds that figure by a substantial amount.

In the context of *Northshore School District*, this statistical significance meant that average pay for certified staff and staffing ratio per 1000 students (the independent variables) combined to be a reliable predictor of expenditures per pupil. The F test established that there was a small probability that the calculated coefficient of multiple determination, .75, was due to chance. That coefficient indicated that 75% of the total variation in per pupil expenditure was accounted for by changes in the independent variables.[70] The practical significance of this high calculated F value and high R^2 was that they served as a valuable part of the chain of evidence designed to demonstrate that variations in per pupil expenditures which allegedly

68. *Id.* at 617.
69. *Id.* at 620.
70. 530 P.2d at 186.

denied equality of educational opportunity to children in areas with low property valuation was not due to variations in local tax revenues but to other constitutionally acceptable factors such as numerical differences in population in different regions of the state.[71]

SUMMARY

Statistical methods indicate only whether a particular measurement could have occurred by chance. We are interested in why certain results occur because causation is an issue in many cases and because fact-finders are interested in the reliability of numerical estimates. Statistics do not prove what caused a particular result, though they may eliminate chance as a plausible explanation and may indicate correspondences between factors relevant to a particular alternative causal theory. If a difference or correspondence could have occurred by chance, then the difference is small, the correspondence negligible. A difference or relationship that is likely to have occurred by chance is not statistically significant. Evidence of a high probability of random occurrence may, however, have tremendous practical significance to the parties involved in a lawsuit because it may provide evidence that one of the parties did not cause the outcome of interest. The various statistical tools merely measure different kinds of relationships — differences and correspondences — and test for the probability that the relationship occurred by chance. These tests all involve two key factors, the size of the group involved or the number of categories in the group involved or the number of categories in the group or sample size, and the diversity or variation within that group or sample. Thus the tools and tests are all related to one another. The validity of competing methods depends on the care with which measurements are made and the statistical and legal validity of the factual theories being tested. The validity of measurement techniques is traditionally the statistician's province and the plausibility of factual theories is traditionally the litigator's expertise. Understanding statistical proof is a matter of appreciating the relationship between the two.

71. *Id.* at 192.