Menu

Table Of Contents	CPV: pp. 1-23	CPV: pp. 24-25
CPV: pp. 26-30	CPV: pp. 31-33	CPV: pp. 34-36
CPV: pp. 37-52	CPV: pp. 53-61	CPV: pp. 62-74

Table of Contents

Measures of Significance: Hypothesis Testing

"Sure your patients have 50% fewer cavities. That's because they have 50% fewer teeth!"

Where Are All of Those Tests?

Hypothesis testing is a widely used approach to actual research problems. And yet, this excerpt may represent the first contact with formal statistical decision making for many readers. If statistical decisions in government, industry, and science have indeed had such a profound impact on our lives, why haven't we heard of hypothesis testing and statistical decision making before undertaking a systematic study of statistics?

Dr. Irwin D.J. Bross answered this question in the introductory chapter of his delightful book, Design for Decision. He briefly discussed the revolution in many fields of science that occurred with the application of sophisticated statistical methods in the period from 1920 to 1940. Then, turning to the matter of public awareness of these methods, he commented.

> I cannot blame you if, at this point, you scratch your head and murmur, "All this looks suspiciously like the old ballyhoo. If Statistical Decision is such a world-shaking affair why haven't I felt some of the tremors?" You may not have heard of the statistical "revolution" that I mentioned earlier, and, to digress a bit, let me explain why you may not have heard of these matters. The main reason is that publications on the subject are written only for fellow specialists (and even these worthies have trouble understanding them). It may take twenty years before these ideas reach other scientists in a comprehensible form and even longer before they are taught to students. Specific techniques (in cookbook form) may be transmitted more rapidly, but the ideas diffuse very slowly.
>
> A few scientists, it is true, have tried to write for the public. But while the public has eagerly accepted the television sets, wonder drugs, and bigger strawberries that scientific research has produced, they have been profoundly uninterested in the fundamental ideas, the Scientific Method, that have made this research fruitful. People must have the very latest electronic gadget, but they cling tenaciously to ideas and methods of thinking that were obsolete three hundred years ago.
>
> This delay in the transmission of ideas is, I believe, one of the factors which has led our civilization to its present crisis. Moreover, the already dangerous situation is steadily getting worse because it is increasingly difficult to translate the language of science — a symbolic one — into everyday English.[1]

Take another look at the way scientific results are reported in the magazines and newspapers. Occasionally you may find reference to probability, but rarely, if ever, will you see a reference to a hypothesis test or to a statistical decision. Rejected null hypotheses in the scientific journals have a way of turning into "proof" in the papers.

[1]From I.D.J. Bross, Design for Decision (New York: Macmillan, 1953),pp.3-4.

Understanding and Using Statistics-Basic Concepts, Marty J. Schmidt, pp. 294-295.

University College Library, London.

KARL PEARSON

Karl Pearson (1857–1936), a professor at University College in London, had already published nine books before he turned his abundant energy to statistics in 1893. Of course, Pearson didn't really take up statistics, which was not yet a separate field of study. He took up problems of heredity and evolution, which led him into statistics.

Pearson developed a family of curves—we would call them density curves—for describing biological data that don't follow a normal distribution. He then asked how he could test whether one of these curves actually fit a set of data well. In 1900 he invented a method, the chi-square test. Pearson's chi-square test has the honor of being the oldest inference procedure still in use. It is now most often used for problems somewhat different from the one that motivated Pearson, as we will see in this chapter.

After Pearson, statistics was a field of study. Fisher and Neyman in the 1920s and 1930s would provide much of its present form, but here is what the leading historian of statistics says about the origins:

Before 1900 we see many scientists of different fields developing and using techniques we now recognize as belonging to modern statistics. After 1900 we begin to see identifiable statisticians developing such techniques into a unified logic of empirical science that goes far beyond its component parts. There was no sharp moment of birth; but with Pearson and Yule and the growing numbers of students in Pearson's laboratory, the infant discipline may be said to have arrived.[1]

Chi-Square in Research

The chi-square statistic is very popular in behavioral science research; a glance through any of the psychological or sociological journals, for instance, suggests that a large proportion of the "facts" in these disciplines are based on chi-square evidence. One reason, no doubt, for the popularity of chi-square is that it can be used in many nonparametric tests. This means that the variable under study need not come from a normally distributed population or represent the higher levels of measurement. Following is a short summary of a typical chi-square application from the area of social psychology.

Do Americans support the Bill of Rights? And does it matter who asks them? A number of studies suggest that a majority of adult Americans will not endorse the Bill of Rights (first ten amendments to the U.S. Constitution) when the original text is placed before them but not identified as the Bill of Rights. Dr. William Samuel conducted a study to see if this was true in one area of Sacramento, California, and also whether it made a difference if a "hip" or "straight" canvasser asked for the endorsement.[1]

Thirteen college-age researchers solicited signatures at a number of middle-class homes in a Sacramento suburb. Seven were dressed in "straight" attire and six in "hip" costume. Each researcher carried three different statements: One was a paraphrased version of the real Bill of Rights (which guarantees a number of basic freedoms), one was a negative paraphrased version that urged restriction of these rights, and a third was a "wishy washy" paraphrase that attempted to take a middle ground between the other two versions. At each house a researcher introduced himself or herself as a representative of a student group called Youth for America; the resident was then asked to read one of the paraphrases and to sign it if he agreed with it.

There was thus two independent variables: attire of the canvasser ("hip" or "straight") and version of the Bill of Rights (real, negative, and "wishy washy"). All respondents were exposed to one level of each independent variable, and all were measured on the same dependent variable, "signature or no signature".

As Samuel expected, people approached on three days of testing[2] were apparently more ready to sign the negative version than the real or "wishy washy" version: 62 percent of those approached endorsed the negative version, while only 46 percent signed the "wishy washy" text, and 44 percent the real paraphrase. But are these differences significant? Is the difference between 62 percent and 44 percent, for instance, due to the effect of the independent variable "version read," or to chance variability. When the chi-square was computed for the appropriate frequencies, the difference appeared to be real ($X^2 = 7.52$, $df = 2$, $p < 0.025$). Thus, the chi-square test supports the conclusion that most of these people would not endorse the Bill of Rights.

However, some other interesting results appeared on closer examination of the data. The above-mentioned preference for the negative version seemed to hold only when the canvasser was dressed as a "straight"; with "hip" canvassers, the frequencies of signatures for different versions was non-significant ($X^2 = 1.46$, $df = 2$). Furthermore, "straights" were more likely to obtain signature than "hips" (for "hip" and "straight" signature rates, $X^2 = 4.46$, $df = 1$, $p < 0.05$).

[1]W. Samuel, "Response to Bill of Rights Paraphrases as Influenced by the Hip or Straight Attire of the Opinion Solicitor," Journal of Applied Social Psychology, 2(1972), 47-62.

[2]The canvassing was actually conducted on four days of testing, but the results were complicated by the fact that one day's canvassing occurred on the Sunday following the Kent State and Jackson State shooting incidents. See Samuel's article for a description of what happened on this day, and the author's theoretical interpretation of these results.

Understanding and Using Statistics-Basic Concepts, Marty J. Schmidt, pp. 367-368.

From the Literature

One of the most famous community studies, *Elmtown's Youth: The Impact of Social Classes on Adolescents*, was published by August B. Hollingshead in 1949. Hollingshead focused his research on the relationship between the social stratification system and the social behavior of adolescents (aged 13 to 19) in Elmtown. Elmtown was studied because it was thought to be a "typical Middle Western community" with a clearcut class structure: upper, upper middle, middle, lower middle, and lower.

Hollingshead was particularly interested in showing the dependence of a series of social behaviors on social class position (the independent variable). He used chi-square, a simple, effective statistical tool, to test the dependence of particular social behaviors on class position.

Because so few of the adolescents were from upper-class families, he combined the upper and upper-middle classes into a single category. Thus the independent variable, social class, was divided into four categories: upper and upper-middle, middle, lower-middle, and lower. The number of categories for the dependent variables ranged from two (working mothers versus nonworking mothers, for example) to six (religious affiliation: Federated, Methodist, Lutheran, Catholic, Baptist, and no affiliation).

Some of the major findings from this study, including the chi-square test results and p-values, follow:

> The higher the class, the more likely parents were counseled on the schoolwork of the child; the lower the class, the more likely parents were counseled on the discipline of the child ($\chi^2 = 19.41, p < .01$).
>
> The higher the class, the greater the number of athletic events, high school dances, and evening plays and parties attended ($\chi^2 = 152.91, p < .01$, for athletic events; $\chi^2 = 95.41$, $p < .01$, for dances; and $\chi^2 = 131.24$, $p < .01$, for plays and parties).
>
> Church affiliation is related to class position: Higher classes are affiliated with the Federated and Methodist churches, the lower classes with the Lutheran, Catholic, and Baptist churches ($\chi^2 = 300.00, p < .01$).

Lyman Ott, Richard F. Larson, & William Mendenhall, Statistics (Boston: Duxbury Press), 1987: 377.

WILLIAM S. GOSSET

What would cause the head brewer of the famous Guinness brewery in Dublin, Ireland, not only to use statistics but to invent new statistical methods? The search for better beer, of course.

William S. Gosset (1876–1937), fresh from Oxford University, joined Guinness as a brewer in 1899. He soon became involved in experiments and in statistics to understand the data from these experiments. What are the best varieties of barley and hops for brewing? How should they be grown, dried, and stored? The results of the field experiments, as you can guess, varied. Statistical inference can uncover the pattern behind the variation. The statistical methods available at the turn of the century ended with a version of the z test for means—even confidence intervals were not yet available.

Gosset faced in his job the problem we noted in using the z test to introduce the reasoning of statistical tests: he didn't know the population standard deviation σ. What is more, field experiments give only small numbers of observations. Just replacing σ by s in the z statistic and calling the result roughly normal wasn't accurate enough. So Gosset asked the key question, What is the exact sampling distribution of the statistic $(\bar{x} - \mu)/s$?

By 1907 Gosset was brewer-in-charge of Guiness's experimental brewery. He also had the answer to his question and had calculated a table of critical values for his new distribution. We call it the t distribution. The new t test identified the best barley variety, and Guinness promptly bought up all the available seed. Guinness allowed Gosset to publish his discoveries, but not under his own name. He used the name "Student," and the t test is sometimes called "Student's t" in his honor. Gosset's statistical work helped him become head brewer, a more interesting title than professor of statistics.

The t-test for the significance of the difference between two means has been applied to questions in all areas of science, from agronomy to zymurgy. Here is one example from psychology.

Do noisy environments affect human performance? Some people are required to spend their working hours amidst high noise levels. Those who live near airports, railroad tracks, and busy highways may be similarly condemned throughout their leisure hours. It is fairly well established that long exposure to even moderate noise levels leads to some permanent hearing loss. Researchers are less in agreement, however, about the effect of noise on human performance in different situations.

Finkelman and Glass reasoned that predictability of noise might be an important factor in determining whether or not performance is impaired by noise when people are working at the limits of their mental ability. To test this idea, they had volunteer subjects carry on mental tasks while being subjected to noise. In the repeated-measures experiment, subjects receiving the "predictable noise" condition were treated to blasts of 80-db noise through earphones, presented for 9 sec. each time, with blasts spaced at regular intervals. Under the "unpredictable noise" condition, subjects received 80-db noise at irregular intervals, in blasts of varying duration. All subjects were scored on a dependent variable "number of errors on a digit recall task."

The mean number of errors on the task was 4.0 for the "predictable noise" group and 8.0 for the "unpredictable noise" group. The obtained t for this difference was 2.37, significant at the 0.05 level (degrees of freedom were unspecified). The obtained significant difference suggests that some kinds of noise conditions do affect performance.

[1]J.M. Finkelman and D.C. Glass, "Reappraisal of the Relationship Between Noise and Human Performance by Means of a Subsidiary Task Measure," Journal of Applied Psychology 54(1970): 211-213.

Analysis of Variance in Research

Without question, the analysis of variance is the most popular statistical technique applied to controlled experiments in the behavioral sciences. A conceptual understanding of ANOVA principles is necessary if one hopes to comprehend and evaluate much of the current research literature in psychology, sociology, and education. An introduction to ANOVA, therefore, is an important part of a first course in applied statistics, even though very few students actually go on to perform experiments themselves.

A quick glance through publications such as the Journal of Experimental Psychology, Sociometry, the Journal of Applied Psychology, or any of the other current behavioral science journals will reveal that ANOVA results are not always presented in the same form. Frequently, you will find ANOVA tables that are abbreviated forms of the tables illustrated in the course. For instance, in reporting the results of an experiment designed to examine the effects of "Speaker Credibility" on "Persuasion," the experimenter might tell you only that "The effect associated with credibility was significant ($F = 34.7$, $df = 3.76$, $p < 0.001$)." Consider for a moment the information contained between the parentheses, and you will recognize that it contains all of the important information displayed in larger tables with SS, MS, df, and "Source of Variation" columns. Because you know that there are three degrees of freedom associated with the F-obtained numerator variance, you will infer there were four experimental groups, each of which received different levels of an independent variable called "Credibility." The error variance has 76 degrees of freedom associated with it. Without even reading the "Methods" section of the article, you should be able to calculate that each group had twenty subjects. (df error $= k(n-1)$). If df error $= 76$ and $k = 4$, then n must equal 20. What about the individual SS and MS values? It is true that you cannot recover them from the parenthetical information, but then it is unlikely that you would ever want to. It is the ratio of the two MS values as expressed by F-observed that helps you evaluate the experiment, not the separate SS and MS values. In the last decade, authors and editors have begun to realize that printing complete ANOVA summary tables is a waste of space, and current reports are likely to omit all but the essential information.

Understanding and Using Statistics-Basic Concepts, Marty J. Schmidt, pp. 393-394.

Does Party-Going Increase Smoking Behavior?

Many cigarette smokers report that they smoke more at parties or other social gatherings than they do otherwise. Is this true? If so, why?

A recent pair of studies by Brett Silverstein, Lynn Kozlowski and Stanley Schachter was conducted to address some of these questions.[1] In an earlier study, these researchers had demonstrated that manipulation of urinary pH (acidity) can influence the number of cigarettes people choose to smoke. In the studies discussed below, they investigated the possibility that some aspect of social situations raises urinary acidity, thereby producing more smoking.

In one study, eighteen smokers simply recorded the number of cigarettes smoked during the day; from these data, the mean number of cigarettes consumed each waking hour was determined for each person. Subjects also kept track of their social activities, and each smoker's days were individually categorized as "social" or "nonsocial", according to several criteria. For these subjects, the mean number of cigarettes smoked during the social day was 31.23, and the mean number during the nonsocial day was 27.85, a significant difference ($t = 2.71$, $df = 17$, $p < 0.02$). Because it was arguable that this difference arises from the simple fact that social days were generally longer than nonsocial days, the mean number of cigarettes smoked per hour on these days was also determined. On social days, subjects averaged 1.85 cigarettes an hour, and nonsocial days they averaged 1.73. This difference between means was almost, but not quite, significant at the 0.05 level ($t = 2.01$, $df = 17$, $p < 0.06$). From this result, the experimenters concluded that these people did smoke more when engaged in social activity. The next step was to examine the role of urinary pH in this situation.

For smokers, it was determined that urinary pH readings were lower at the end of social days (mean = 5.86) than at the end of nonsocial days (mean = 6.30); this difference was statistically significant ($t = 3.49$, $df = 15$, $p < 0.01$). But does this drop cause more cigarette smoking? In further tests with other subjects, pH reading dropped from a mean of 6.43 before a two-hour party to 6.00 after the party ($t = 3.23$, $df = 14$, $p < 0.01$). Moreover, this drop appeared for both smokers and nonsmokers, refuting the possible argument that smoking causes pH changes rather than the reverse. These results, along with the earlier evidence that pH manipulation alters smoking behavior, suggest a partial explanation for increased smoking at parties.

[1] B. Silverstein, L. Kozlowski, and S. Schachter, "Social Life, Cigarette Smoking, and Urinary pH," Journal of Experimental Psychology: General, 106 (1977): 20-23.

Understanding and Using Statistics-Basic Concepts, Marty J. Schmidt, pp. 320-321.