Research Methodology: Some Statistical Considerations

Peter D. Hurd

ABSTRACT: This article is the third in a series of three designed to help managed care pharmacists design and interpret research studies. The objective of this article is to provide a general introduction to the use of statistics in research articles. In the changing health care industry, an overview of basic

statistical concepts will benefit the pharmacist practicing in a managed care environment.

Key Words: Statistics, Variation, Significance, Statistical tests

J Managed Care Pharm 1998; 617-621.

This article is the third in a series of articles about research methodology^{1,2} written for those who are relatively unfamiliar with statistics and research in managed care. It is designed to provide a broad overview of statistical issues. This article cannot serve the place of a textbook or course on statistics, but it can provide some of the tools needed to understand and use research articles while avoiding some of the common misinterpretations of research findings. Those who have not read the previous two articles in the series can benefit from this article alone, but reading all three in sequence would be the best approach.

VARIATION

One of the central concepts in statistics is variation. When comparing two groups—for instance, one group that has received a placebo and another that has received an active drug—the responses of the individuals in each group will show variations. Some variation will be due to the difference between the placebo and the drug, but additional variation will be due to the differences between the people receiving the placebo (perhaps taller, or older, or sicker, or less liver-impaired) when compared to those who received the active drug. How can we tell if the difference between the two groups is due to the effect of the drug or to the effects of the variation in the people assigned to the two groups?

CLINICAL AND STATISTICAL SIGNIFICANCE

Statistics provide one of many ways to detect differences between groups. Sometimes the variation in responses of one group is so great in comparison to the other group that the difference obviously exists even without calculating the statistics. For example, assume that advertising begins for a drug information 800-number. Immediately after the advertising begins, calls increase from two per day to more than 80 per day. Statistics are unnecessary to reveal things already known, such as that the advertising has been successful; the increase in calls is significant; and more people will be needed on the

AUTHOR

PETER D. HURD, PH.D., is Professor and Director of the Division of Administrative Sciences and Liberal Arts at the St. Louis College of Pharmacy, St. Louis, MO.

AUTHOR CORRESPONDENCE: Peter D. Hurd, Ph.D., St. Louis College of Pharmacy, 4588 Parkview Place, St. Louis, MO, 63110.

ACKNOWLEDGEMENTS: The author would like to thank Brenda R. Motheral, R.Ph., Ph.D., Assistant Professor at the University of Arizona College of Pharmacy, and Kenneth W. Schafermeyer, Ph.D., Associate Professor and Director of Graduate Studies at the St. Louis College of Pharmacy for their comments on an earlier version of the manuscript.

CE CREDIT: This is article number 233-000-98-006-H04 in AMCP's continuing education program. It affords 1.0 hours (0.01 CEU) of credit. • Learning objectives and test questions follow on page 622.

phones. At other times, only one case can detect a difference. For example, if a patient's cancer therapy is so expensive that the drug budget for the year is spent, it's clear that this year's drug budget will be significantly higher than the previous year's.

Sometimes, however, the variation between groups is not clearly different from the variation that exists within each of the groups. Statistical significance provides an additional means of asserting that the variation between situation one and situation two is great enough to conclude that they are different. To do this, we rely on the fact that our world has certain patterns of variation. For example, consider coin tosses. We expect that heads will appear half the time and tails the other half with an unbiased coin, and we can extrapolate that expectation to the results of tossing the same coin a number of times or tossing a lot of coins all at once. If, over the long run, the results of tossing the coin vary too much from 50/50, then the coin probably is biased. Statistics sets the rules for what is "too" different. A similar example would be the normal distribution, or the bell-shaped curve. Some things in nature show a variation that resembles this curve; some faculty members feel that performance on a test should fit this distribution. Again, statistics sets the rules to help a researcher determine if the distribution of scores from one group (the active drug group) is different enough from the distribution of scores from another group (the placebo group) to conclude that this much variation would not be expected if the two groups were responding in the same way.

What does it mean to say that a difference is statistically significant? In very general terms, this means that the findings are unusual enough that one could not reasonably expect that the difference is due to chance or expected differences between groups that result from random variation, measurement error, and less-than-perfect research design. The amount of variation between the two groups usually would not be observed if both groups really represented the same underlying distribution of responses.

In a nutshell, the researcher is saying that if the study were repeated, there is a strong likelihood that the findings would be the same; the difference would be found again.

This brings us to something called the *p*-value, most often seen in a research article, expressed as $p=0.05$ or $p<0.05$. Think of "*p*" as representing "probability." This means that the researcher, using some statistical test, has determined that the chance of random variation causing these differences is less than five in 100. By tradition, when random variation is this unlikely to have caused a difference ($p<0.05$), that difference is considered statistically significant.

Statistical differences have a margin for error—a fact that may be troubling to some. Two types of errors have special importance. There is a chance that a researcher may conclude that two groups are different, when in fact if the study were repeated, the results would show that the groups are not really different. In statistics, the risk that this will happen five times in 100 ($p<0.05$) is acceptable. This type of error, in which a

statistical difference is attributable to a mistake, is called a Type I (or Type One) error. A Type II (Type Two) error occurs when the researcher concludes that there was no difference when in fact a difference actually existed. A Type I error can be compared to a diagnostic test that gave a false positive result and a Type II error to a false negative, in which the test failed to detect a problem that the patient really had. So statistical conclusions can be made in error, just like other types of conclusions.

Another caution: Because of the nature of statistics, differences that are statistically significant may have less value in terms of clinical significance. Although the same difference would be expected again from a statistical standpoint, one also should ask if this difference is meaningful or important.

THE NUMBER OF SUBJECTS AND SIGNIFICANCE

The discussion of statistical significance raises an issue about the number of subjects, patients, clients, or members in a research design. Which is better: a large number of subjects or just a few subjects? With just a few subjects, a researcher would be unable to detect rare side effects of a drug. With a large number of subjects, research could be wasting funds if it found a significant difference with the first 15 patients in a group, but continued to run hundreds more subjects through the study only to find the same difference. The answer, of course, is that the number of subjects depends on the situation and the research question that one is attempting to answer.

As a general rule, the more subjects used in a research project, the easier it is to find a statistically significant difference. This does, however, raise the issue of statistical versus clinical significance. For example, a study might find a compliance rate of 77% among patients taking a drug once a day, whereas the compliance rate of patients taking the same dose twice a day is 75%. With a large number of subjects, this difference could be deemed statistically significant. Yet this difference may have little, if any, clinical meaning.

Nonresponse bias is another problem associated with the number of subjects and the interpretation of significance. In this case, we usually are talking about a survey that is answered by a percentage of those who have received the survey. As a general rule, response rates greater than 50% are less likely to show response bias than those with response rates of less than 50%. The logic is very straightforward: If less than half the people respond, then it is possible that they feel totally different from those who did respond. If more than half respond, the study is at least dealing with a majority of the possible responses. If the survey design encourages some to respond and others not to respond (for example, only those who have a problem with a service tended to respond), then the researcher could draw conclusions that a wider-ranging survey with a higher percentage of response would not support. Studies with a low response rate should be treated with considerable caution.

Studies with a low response rate often will attempt to rule out biases by an examination of their data. For instance, data from early responders will be compared to the responses of the last individuals who reply to a survey. When the responses are similar, there is less chance for bias. For example, a study of pharmacies in the United States could show that, despite a low response rate, the percentage of busy stores that returned the questionnaire early was similar to the percentage that returned the survey late. In a survey that showed the later responders tended to be the busier stores, one might suspect a potential bias based on how many prescriptions were filled per week. When the potentially biased study demonstrates significant differences between groups, this tendency for the busiest stores to be underrepresented needs to be considered in the interpretation of the data.

Another way to examine the potential for bias in a survey with a lower response rate would be to compare the demographics of those who responded to the demographics of the entire population. In a national survey of patients in a particular managed care plan, a researcher could compare the geographic distribution, ages, gender mix, and health problems to that of the total care plan. When the respondents' characteristics are similar to those of the total population, there is less chance that these factors will bias the results.

When studying patients who have one or more diseases, researchers must consider the availability of appropriate participants. One of the classic problems encountered in clinical research is a study design that requires more patients than the clinic normally would see during the time allotted for the study. One way to examine this problem is to study the history of the patient population and make a very conservative estimate of the availability of appropriate participants, remembering to apply the exclusion criteria planned for the study to determine the number of subjects for the project. Studies that are unable to enroll enough subjects may lack the statistical power to find a difference when one exists (a Type II error). For example, studies that report that two medications showed no treatment differences but rely on very small samples (perhaps about 10 patients per treatment) simply may have failed to detect a more subtle difference because of the limited number of subjects.

PARAMETRIC AND NONPARAMETRIC STATISTICAL TESTS

As described in the first article of this series, numbers are used in many different ways. Sometimes numbers are used as a shorthand name: group 1 and group 2, or baseball players' numbers on their uniforms. The baseball player using "16" is not half as good as the player using number "32"; these are just identifiers. This use of numbers is called "nominal" because there is little mathematical meaning to the numbers, just a categorization.

In other instances, a number signifies a particular order, but nothing else. Ranking the best hospitals in the country as

first, second, third, and fourth should not imply that number two is twice as good as number four. But two has been judged more favorably than four, and this difference is presented as if it is meaningful. This is an ordinal use of numbers because the numbers' order has meaning.

Sometimes the numbers' intervals have a very real meaning. A temperature of 50 degrees Fahrenheit is in fact 10 units less than a temperature of 60 degrees. In this case, though, a 0-degree temperature does not imply the absence of any temperature. The use of numbers when the interval is assumed to be equal across the scale, but zero does not mean the absence of the property, is called interval data.

Often the data has a zero that means absence of a condition—no heart rate, zero on a test, or zero money in the checkbook. This use of numbers is called ratio data. In managed care, this is often the type of data analyzed: number of medications, number of side effects, cost of the procedure, and so on.

Statistics have been devised for all four types of data. Usually nominal and ordinal data are analyzed by nonparametric statistics (e.g., chi square is one of the most familiar nonparametric tests), and interval/ratio data usually are analyzed with parametric statistics (e.g., t-test, analysis of variance, regression).

The mean (the sum of the numbers divided by the number of the data points) helps illustrate this. Will the mean make sense with nominal data? No. Why would anyone ever add up the numbers on the team's roster and then divide by the number of players? This would not provide useful information. Similarly, would it make sense to add up the rankings of the different hospitals in each state to come up with an "average" rank? Well, this is done, but ordinal data's lack of equal intervals makes these numbers difficult to interpret. The median response (the response in the middle of the responses after they have been put in order) or the modal response (the response used most often) would be more appropriate for ordinal data.

In the following sequence, what is the modal response? The median response?

1, 1, 1, 2, 4, 5

The modal response would be 1; the median response would be 1.5 (halfway between 1 and 2, because there are an even number of responses); and the mean would be $14/6$ (which is 2.33 if one compares the three measures). Each of these has a use, but the mean response is most commonly reported.

The use of the median can help when outliers might distort the interpretation of the mean of a set of numbers. The housing market is a good example. Although the mean price of a house in a community is useful information, a few very expensive houses will distort this number so that it is much higher than the price most people would pay for homes in that area. Consequently, the median response often is used in local housing reports.

Much as the mean is most appropriate for interval and ratio data, two measures of variation also are based on equal

intervals between numbers: variance and standard deviation. Variance refers to the amount of variation around the mean of the data being analyzed. As a general rule, a larger variance would indicate more variation in the data. Conversely, if everyone chose the same response, then there would be no variance in the responses.

Here is an illustration of the importance of considering both mean and variance rather than only the mean. It is possible on a survey to have a mean of 3 on a 5-point scale (1–2–3–4–5) with no variance—everyone chose 3. If half the respondents chose 1 and the other half chose 5, then the mean would still be 3, but the variance would be quite large, such as the difference between a barbell and a rolling pin.

xxx-----xxx

---xxxxxx---

Based on this example, it should be clear that both mean and variance are important in the analysis of this type of data.

Technically, standard deviation is the square root of the variance. From a more practical standpoint, the standard deviation is easier to compare with other standard deviations and is usually presented in research articles. Without getting very specific, the variance is computed by squaring numbers, whereas the standard deviation is more applicable to the particular measuring scale that was used. (The term "standard error" is not the same and may be smaller than the standard deviation.)

CORRELATIONS

The association between two variables sometimes is analyzed using correlations. Some tests for correlation are parametric (e.g., Pearson product moment correlation coefficient), and some correlation tests have been designed for ordinal data (e.g., Spearman rank correlation). The fundamental question being asked has to do with the relationship between the observations. For example, as age increases does drug use increase? Correlations provide a statistical way to decide if the relationship between observations could be the result of chance or if there is some pattern that could not be expected if the two observations were unrelated. As mentioned above, the more subjects in the analysis, the easier it is to find statistically significant correlations, but they may be of little clinical interest.

Correlations can be either positive or negative. Positive correlations mean that a larger response on one variable is associated with a larger response on the second variable. In pharmacy, for example, the more medications a patient is taking, the more side effects the patient is likely to experience. This would be a positive correlation between the number of medications and the number of side effects. A negative correlation means that a higher score on one variable is associated with a lower score on the second variable—for example, usually the more medications individuals are taking, the poorer their health would be.

Is the following a positive or negative correlation? The faster the car goes, the poorer the gas mileage. Correlating the faster speed with fewer miles per gallon gives a negative correlation. Sometimes, however, one could associate the faster speed with more use of gas, a positive correlation. So, it depends on what is being measured and how the data are reported.

The correlation of two variables does not necessarily mean that one is the cause of another. Correlation does not equal cause. For example, one might find that the sale of ice cream increases with the number of drownings. Does the ice cream cause the drownings? Probably not. Both are associated with another variable—warm weather. People eat more ice cream and participate in more water sports in hotter weather. Avoid the error of assuming cause when all that has been demonstrated is correlation.

A number of statistical tests are related to correlation. For example, regression is an attempt to fit a line through a distribution of data. That line would be similar to a representation of the correlation between the two variables. Multiple regression is an elaboration of simple regression, using a number of variables to discover the best combination of variables to predict the targeted variable. Factor analysis also is based on association between variables. In this case, the statistical procedure is looking for various factors that seem to be related. Although factor analysis and multiple regression are complicated statistics that rely on computers for calculation, each tries to determine the association between different variables that the researcher is studying.

DIFFERENCES BETWEEN GROUPS

Sometimes a researcher wants to determine whether one group is different from another. For example, have subjects taking the active drug responded better than those in the placebo group? There are statistical tests to help answer this question; some are parametric and others are for nonparametric data. This section will concentrate on two parametric tests: the t-test and the analysis of variance (ANOVA).

The t-test is designed to detect statistically significant differences between two groups. For this test to work, it assumes that there are no systematic biases between the groups, a circumstance often accomplished by random assignment to conditions and good data collection techniques. Of course, in the case of clinical trials, there is one intentional difference between the groups: half received the placebo and half received the active drug. To analyze the data the researcher needs each subject's results and to know if they received placebo or active drug. This information is entered into the computer, and a t-value is generated. A larger t-value is more likely to have a p-value that is less than 0.05. When the t-value is large enough to have a p-value less than 0.05, the researcher will conclude that the difference is statistically significant. In our

terms, that indicates that the variation between the two groups was too great to happen by chance (at least 95 times out of 100, anyway). We feel confident saying that the groups have significantly different values for the dependent variable being measured.

Sometimes the difference between two groups is based on a study design that uses the same individual or matched individuals—for example, before and after taking a medication. This is a special case of the t-test because the two groups are not independent. The t-test is modified statistically to account for this expected change in variation and reports a matched, paired, or dependent t-test value. Again, the larger the t-value, the more likely the researcher will be able to say that the p-value is less than 0.05, and the groups appear to be different.

This use of the words “independent” and “dependent” refers to the subjects analyzed with the t-test. These words have been used in a different way when referring to dependent and independent variables. As described in the first article in this sequence, dependent variables are those variables that one is trying to explain. In a clinical trial, the dependent variable would be the patient’s response to either the treatment or placebo medication. The independent variables are those manipulated by the experimenter (e.g., random assignment to receive drug or placebo) or expected to affect the dependent variable (e.g., in the ice cream and drownings example, the ice cream was presented as the independent variable). The use of the term “independent variable” is slightly different for experiments (drug/placebo) and for correlations (drownings and ice cream) as illustrated above. Considering the context in which the term is used will help clarify what the researcher is describing.

The t-test is limited to two groups, whereas ANOVA is designed to handle more complicated research designs. For example, a researcher would use ANOVA to determine if any of three conditions differs from the others, as when comparing a low dose, higher dose, and highest dose of a drug. Other types of ANOVA will handle factorial designs, such as a 2x2 experiment in which patients are randomly assigned to one of four conditions—receiving low or high doses of a drug and low or high levels of follow-up. This results in four cells: low/low, low/high, high/low, and high/high. ANOVA also will be able to handle more complicated factorial designs (e.g., 2x3x2). Again, the independence of subjects in each cell is assumed, usually through randomization. Collecting data from subjects in a pre/post-test design requires a modification of the statistical test to account for this fact.

This very brief description of these tests is meant only as an overview. It is hoped that a general understanding of the terms and the concepts will establish a foundation for further study and will assist the reader in the literature that uses these approaches.

QUALITATIVE DATA

Because this article deals with numbers and statistics, little has been said about the value of qualitative data. Qualitative data can range from the question at the end of a survey that asks for any comments that the individual would like to provide, all the way to an elaborate collection of stories and anecdotes as a part of a sociological analysis of a company. Qualitative information can be extremely powerful; one story of a patient helped by a staff member can carry much more impact than an analysis of patient satisfaction that shows a significant increase from the previous year.

CONCLUSIONS

Perhaps the most important message in the use of statistics is to plan statistical analysis at the very start of the study design. Perhaps the researcher should determine the type of analysis that will be done, or consult with a statistics expert before collecting the data. After data collection has started, it can be expensive or impossible to repair faulty design.

The first article in this series discussed the development of hypotheses. Hypotheses have a direct link with the statistical analysis of a project. Being clear and precise about both will greatly aid the researcher as the project develops. Both should be formalized before the project moves to the data collection stage. The second article discussed research methodology. The planning of the statistical analysis should be integrated with the design of the research project. The sequential nature of these articles does not necessarily imply an order of events that must take place in the implementation of a research project.

Numbers need to be interpreted in the context of the patients that they represent. The numbers are abstractions of real people, suffering real pain and providing real answers. A good researcher tries to use the numbers to summarize the responses of a group of people, while remembering that a very real heartbeat is associated with each of the responses.

▲ References

1. Motheral BR. Research methodology: hypotheses, measurement, reliability, and validity. *J Managed Care Pharm* 1998; 4: 382-94.
2. Schafermeyer KW, Hurd PD. Research methodology: designing a research study. *J Managed Care Pharm* 1998; 5: 504-14.

C E E X A M

Research Methodology: Some Statistical Consideration

Upon completion of this article, the successful participant should be able to:

1. explain the term "statistical significance."
2. discuss the impact of the number of subjects on statistics.
3. differentiate between parametric and nonparametric tests.
4. discuss the importance of mean and variance in data analysis.
5. differentiate correlation and causation.

SELF-ASSESSMENT QUESTIONS

1. Statistical tests are:
 - a. either parametric or nonparametric.
 - b. only parametric.
 - c. only nonparametric.
 - d. neither parametric nor nonparametric.
2. When the research article reports $p < 0.05$, this usually refers to a:
 - a. lack of difference between two or more groups.
 - b. statistically significant difference.
 - c. clinical difference that may have little statistical significance.
 - d. small sample size.
3. When the research article reports $p < 0.05$, this would usually mean a ____ t-value.
 - a. small
 - b. insignificant
 - c. large
4. A Type I error would be:
 - a. reporting no difference when one existed.
 - b. reporting a difference when one did NOT exist.
 - c. using too large of a sample size.
 - d. running a parametric statistic inappropriately.
5. Using a t-test with a sample size of 10 to report that two medications were equally effective could be a:
 - a. Type I error.
 - b. Type II error.
 - c. correlation.
 - d. significant difference.
6. A larger sample size ____ a significant difference compared to a smaller sample size.
 - a. makes it harder to show
 - b. has no effect on showing
 - c. makes it easier to show
7. If the variance of a sample was four, which number would be larger?
 - a. Variance
 - b. Standard deviation
8. When two variables are correlated, then one can be said, statistically, to cause the other.
 - a. True
 - b. False
9. When two means from two questions on a survey are about the same:
 - a. the standard deviation of the scores will be the same.
 - b. the distribution of the scores will be different.
 - c. the standard deviation of the scores might help understand the results.
 - d. the data cannot be analyzed with statistics.
10. To test for significant differences between three groups, the best choice is:
 - a. correlation.
 - b. t-test.
 - c. analysis of variance.

See text of article beginning on page 617 of this issue of *JMCP*.

This article qualifies for 1 hours of continuing pharmaceutical education (0.10 CEU). The Academy of Managed Care Pharmacy is approved by the American Council on Pharmaceutical Education as a provider of continuing pharmaceutical education. This is program number 233-000-98-006-H04 in AMCP's educational offerings.