

In search of significant differences

Milton Houpt, DDS, PhD

Consider reading the following: "After lengthy study, no statistically significant differences were found between male and female subjects, however, the *trend* shows . . ." or "Variables 1 and 2 were significantly different at the 0.01 level, but variables 3 and 4 were different at *only* the 0.05 level; however, variables 5 and 6 were different at the 0.001 level."

Such statements indicate a basic misunderstanding of the purpose of statistical testing, and often are prompted by a desire of investigators to find positive differences between groups. The notion that research must demonstrate positive differences between groups in order to be good is mistaken. Careful research should seek to explore *reality* rather than prove a bias.

The following discussion will differentiate briefly between descriptive and inferential statistics in order to consider the purpose of statistical testing. Two guidelines for investigators are presented. When statistical tests are to be performed, the level of probability should be determined *prior* to the analysis of that data. More important than statistical testing is the interpretation of findings to determine if statistically significant differences are *meaningful* within the context of the research questions being asked.

Descriptive and Inferential Statistics

Measurements made on a group of subjects are often quite voluminous. Consequently, they usually are summarized and represented by an *average score* (usually the arithmetic mean) and some *measure of variability* (usually the standard deviation), that is, how different all other scores are from the average score. In a large group most variables are distributed with most subjects clustering around the mean so that roughly two-thirds of the group are within ± 1 standard deviation from the mean. For example, in regard to height for a particular group, most subjects were an average height of 5'8". If the standard deviation was 2", approximately two-thirds of the group would

be between 5'6" and 5'10" and approximately 95% of the group would be between 5'4" and 6' in height. The mean and standard deviation are descriptive statistics for the complete group being studied.

If the total group being studied is very large, it may not be desirable or even possible to collect data from all members of the group. Rather than the total population, a smaller *sample* is studied with the assumption that the sample is a fair representation of the population. Purely by chance the average for the sample might be lower or higher than the average for the population. The greater the size of the sample and/or the smaller the variability of the population, the closer will the sample average approximate the population average (Fig 1). Samples are used to make inferences about the total population (hence inferential versus descriptive statistics).

If 2 samples are studied, they may represent 1 population or they may represent 2 different populations. Inferential statistics are used to make that determination.

Testing for Differences Between Groups

When an investigator studies and compares 2 groups, it is to determine if those samples represent 2 different populations or whether they are derived from the same population (Fig 2). Statistical tests are used in order to determine if there are *statistically significant differences* between 2 groups. Conceptually, this involves comparing the *variability within* the groups to the *variability between* the groups. If there is little variability within the groups and great variability between the groups, chances are the groups will represent 2 different populations (Fig 3A). In the reverse situation, chances are the groups will be similar and represent 1 population (Fig 3B). True differences between groups may or may not exist; that is, the 2 groups may represent 2 different populations or 2 samples of the same population. Differences may

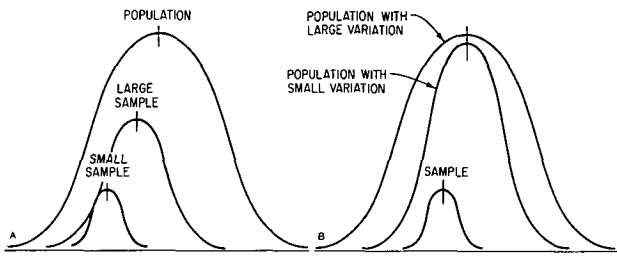


FIG 1. A. (left) Frequency distributions for a population and samples which are small and large. Note that the larger sample has a mean closer to the population mean than the smaller sample. B. (right) Frequency distributions for populations with small and large variability and a sample. Note that the sample mean more closely approximates the mean of the population with the smaller variability.

be due to chance sampling variation from 1 population, or there may be real differences between the 2 populations (Fig 2).

The truth about groups cannot be known with absolute certainty. Rather, only inferences from observed data can be made, assuming that the data are a fairly good representation of the true values. Because the truth cannot be known, the investigator may produce an error. The investigator might conclude that there were real differences between groups when, in fact, there were really no differences other than those caused by chance sampling variation. This would have been a *false positive* finding referred to as a Type I error. On the other hand, the investigator might suggest that there were no real differences between the groups when, in fact, there were differences other than those caused by chance. This would have been a *false negative* finding and a Type II error would have been committed.

Type I and Type II errors are inversely related; that is, as the probability of making 1 type of error is decreased the probability of making the other type is increased. The researcher must select the level of probability in order to use statistical tests. The researcher must decide how much chance variation between groups can be accepted before a judgment of true difference can be made, or rather how much Type I compared with Type II error can be accepted.

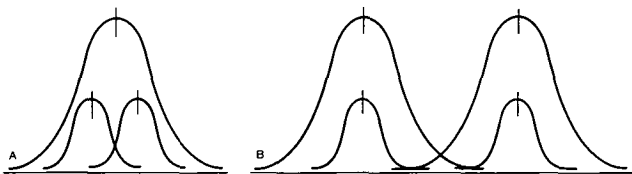


FIG 2. A. (left) Frequency distributions for 2 samples of 1 population. B. (right) Frequency distributions for 2 samples of 2 different populations.

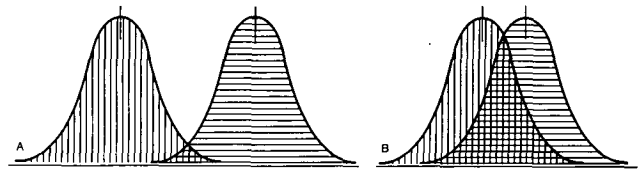


FIG 3. A. (left) Two frequency distributions with great between group variation. Note that there is little overlap of the 2 distributions. B. (right) Two frequency distributions with little between group variation. Note that there is a great amount of overlap of the 2 distributions. Fig 3A would probably represent 2 samples taken from 2 different populations, as compared with Fig 3B which would represent 2 samples taken from the same population.

Earlier it was stated that 95% of the means of random samples would be within ± 2 standard deviations of the population mean and 5% of sample means would be more than 2 standard deviations from the population mean 5% of the time. If the 95% probability level was used and a sample mean was found to be more than 2 standard deviations from the population mean, the researcher would conclude that the sample was probably from a different population (even though that mean would be found 5% of the time purely by chance).

In dentistry, a 95% probability level frequently is used. In psychology, where measurement might not be as precise, the 90% level sometimes is used, whereas in materials research where instruments and techniques are extremely accurate, the 99% to 99.5% level is employed. The level of probability is determined not only by the accuracy of measurement, but also by the importance of the questions being asked. If the experiment was performed to differentiate 2 new metals, 1 of which was to be used in a new airplane in which the investigator was to be the first passenger, then it is quite likely that a very high level of probability would be chosen!

Since both Type I and Type II errors can be made, it becomes important to choose the level of probability *prior* to the performance of the experiment. When the data are analyzed and tested statistically, a determination of difference or no difference between groups can be made. If no statistically significant difference is found, it should be reported as such because the finding of no difference could be just as important as a finding of a difference between groups. In sound research the investigator should devote efforts to conducting an unbiased experiment, rather than attempt to manipulate the data to show positive differences.

Meaningful Versus Statistically Significant Differences

Once differences between groups have been determined to be statistically significant, there yet remains the question as to whether those differences are meaningful. This becomes a rather arbitrary but important interpretation of the findings with specific reference to the questions being asked. Consider, for example, the experiment performed in 1969 by the eminent investigator Pop Eye and his associate, Olive Oyle, in order to determine if daily ingestion of spinach appreciably increased the size of the bicep muscle in 10-year-old children.

Following exhaustive data analysis, they determined that there was a statistically significant difference between the spinach and nonspinach eaters in regard to size of bicep muscles. That difference in size was found to be 1 mm. Whether that difference is meaningful becomes an arbitrary decision. Eye and Oyle must have concluded that a difference of 1 mm following 1 year of daily ingestion was not meaningful; however, if that difference developed after *each can* of spinach, they might have concluded that the difference was very meaningful!

In dental research, when statistically significant differences between groups are found, it is important to ask whether those differences are meaningful within the context of the investigation. Should routine culturing be performed in endodontics if that procedure only increased success from 90 to 95%? Should a new caries preventive agent be used if after 4 years the difference between the experimental and control group is only one-fourth of a lesion? Should practitioners be urged to use alloy A which is slightly more costly but lasts twice as long as alloy B?

In order to answer these questions, researchers might look for statistically significant differences between experimental and control groups. Neverthe-

less, it is only through judicious interpretation of the data that meaningful answers to the questions can be found.

Summary

Statistical techniques provide tools whereby characteristics of things can be measured producing data which can be summarized and analyzed. Statistical tests can be used when data from different groups are compared; however, the probability level for the appropriate tests of significance should be determined *prior* to the analysis of the data. If statistically significant differences between groups are found, there remains the more important task of determining if those differences are *meaningful*. Researchers might be surprised and even disappointed when no differences are found between groups, but they should be content with a simple finding that is an addition to our knowledge.

Dr. Houpt is professor and chairman, pediatric dentistry, New Jersey Dental School. Reprint requests should be sent to: Dr. Milton Houpt, Department of Pediatric Dentistry, New Jersey Dental School, University of Medicine and Dentistry of New Jersey, 100 Bergen St., Newark, NJ 07103.

Author's Note: For further elaboration of statistical methods, the reader is referred to an introductory or advanced statistical text, some of which are listed below:

Ary D, Jacobs LC: Introduction to Statistics, Purposes and Procedures. Holt, Rinehart and Winston, 1976.

Chase CI: Elementary Statistical Procedures, 2nd ed. McGraw-Hill, 1976.

Chilton NW: Design and Analysis in Dental and Oral Research, 2nd ed. 1982.

Guilford JP, Fruchter B: Fundamental Statistics in Psychology and Education, 6th ed. McGraw-Hill, 1978.

Weintraub J, Douglass CW, Gillings DB: Biostats, Data Analysis for Dental Health Care Professionals, 2nd ed. Cavco Inc, 1985.

Second time around

Nearly half of the doctors and dentists who responded to magazine surveys said they would choose another field if given a second chance. Forty-one per cent of the 754 doctors and 46% of the 1000 dentists surveyed by *Physicians Management* and *Dental Management* said they would try something else. However, 87% of the doctors and 90% of the dentists said they were satisfied with their incomes.