Statistics: An Introduction

Main Page

Statistics is a set of tools used to organize and analyze data. Data must either be numeric in origin or transformed by researchers into numbers. For instance, statistics could be used to analyze percentage scores English students receive on a grammar test: the percentage scores ranging from 0 to 100 are already in numeric form. Statistics could also be used to analyze grades on an essay by assigning numeric values to the letter grades, e.g., A=4, B=3, C=2, D=1, and F=0.

Employing statistics serves two purposes, (1) description and (2) prediction. Statistics are used to describe the characteristics of groups. These characteristics are referred to as variables. Data is gathered and recorded for each variable. Descriptive statistics can then be used to reveal the distribution of the data in each variable.

Statistics is also frequently used for purposes of prediction. Prediction is based on the concept of generalizability: if enough data is compiled about a particular context (e.g., students studying writing in a specific set of

classrooms), the patterns revealed through analysis of the data collected about that context can be generalized (or predicted to occur in) similar contexts. The prediction of what will happen in a similar context is probabilistic. That is, the researcher is not certain that the same things will happen in other contexts; instead, the researcher can only reasonably expect that the same things will happen.

Prediction is a method employed by individuals throughout daily life. For instance, if writing students begin class every day for the first half of the semester with a five-minute freewriting exercise, then they will likely come to class the first day of the second half of the semester prepared to again freewrite for the first five minutes of class. The students will have made a prediction about the class content based on their previous experiences in the class: Because they began all previous class sessions with freewriting, it would be probable that their next class session will begin the same way. Statistics is used to perform the same function; the difference is that precise probabilities are determined in terms of the percentage chance that an outcome will occur, complete with a range of error. Prediction is a primary goal of inferential statistics.

Revealing Patterns Using Descriptive Statistics

Descriptive statistics, not surprisingly, "describe" data that have been collected. Commonly used descriptive statistics include frequency counts, ranges (high and low scores or values), means, modes, median scores, and standard deviations. Two concepts are essential to understanding descriptive statistics: *variables* and *distributions*.

Variables

Statistics are used to explore numerical data (Levin, 1991). Numerical data are observations which are recorded in the form of numbers (Runyon, 1976). Numbers are variable in nature, which means that quantities vary according to certain factors. For examples, when analyzing the grades on student essays, scores will vary for reasons such as the writing ability of the student,

the students' knowledge of the subject, and so on. In statistics, these reasons are called variables. Variables are divided into three basic categories:

Nominal Variables

Nominal variables classify data into categories. This process involves labeling categories and then counting frequencies of occurrence (Runyon, 1991). A researcher might wish to compare essay grades between male and female students. Tabulations would be compiled using the categories "male" and "female." Sex would be a nominal variable. Note that the categories themselves are not quantified. Maleness or femaleness are not numerical in nature, rather the frequencies of each category results in data that is quantified -- 11 males and 9 females.

Ordinal Variables

Ordinal variables order (or rank) data in terms of degree. Ordinal variables do not establish the numeric difference between data points. They indicate only that one data point is ranked higher or lower than another (Runyon, 1991). For instance, a researcher might want to analyze the letter grades given on student essays. An A would be ranked higher than a B, and a B higher than a C. However, the difference between these data points, the precise distance between an A and a B, is not defined. Letter grades are an example of an ordinal variable.

Interval Variables

Interval variables score data. Thus the order of data is known as well as the precise numeric distance between data points (Runyon, 1991). A researcher might analyze the actual percentage scores of the essays, assuming that percentage scores are given by the instructor. A score of 98 (A) ranks higher than a score of 87 (B), which ranks higher than a score of 72 (C). Not only is the order of these three data points known, but so is the exact distance between them -- 11 percentage points between the first two, 15 percentage points between the second two and 26 percentage points between the first and last data points.

Distributions

A distribution is a graphic representation of data. The line formed by connecting data points is called a frequency distribution. This line may take many shapes. The single most important shape is that of the bell-shaped curve, which characterizes the distribution as "normal." A perfectly normal distribution is only a theoretical ideal. This ideal, however, is an essential ingredient in statistical decision-making (Levin, 1991). A perfectly normal distribution is a mathematical construct which carries with it certain mathematical properties helpful in describing the attributes of the distribution. Although frequency distribution based on actual data points seldom, if ever, completely matches a perfectly normal distribution, a frequency distribution often can approach such a normal curve.

The closer a frequency distribution resembles a normal curve, the more probable that the distribution maintains those same mathematical properties as the normal curve. This is an important factor in describing the characteristics of a frequency distribution. As a frequency distribution approaches a normal curve, generalizations about the data set from which the distribution was derived can be made with greater certainty. And it is this notion of generalizability upon which statistics is founded. It is important to remember that not all frequency distributions approach a normal curve. Some are skewed. When a frequency distribution is skewed, the characteristics inherent to a normal curve no longer apply.

Making Predictions Using Inferential Statistics

Inferential statistics are used to draw conclusions and make predictions based on the descriptions of data. In this section, we explore inferential statistics by using an extended example of experimental studies. Key concepts used in our discussion are probability, populations, and sampling.

Experiments

A typical experimental study involves collecting data on the behaviors, attitudes, or actions of two or more groups and attempting to answer a research question (often called a hypothesis). Based on the analysis of the data, a researcher might then attempt to develop a causal model that can be populations.

A question that might be addressed through experimental research might be "Does grammar-based writing instruction produce better writers than process-based writing instruction?" Because it would be impossible and impractical to observe, interview, survey, etc. all first-year writing students and instructors in classes using one or the other of these instructional approaches, a researcher would study a sample – or a subset – of a population. Sampling – or the creation of this subset of a population – is used by many researchers who desire to make sense of some phenomenon.

To analyze differences in the ability of student writers who are taught in each type of classroom, the researcher would compare the writing performance of the two groups of students.

Dependent Variables

In an experimental study, a variable whose score depends on (or is determined or caused by) another variable is called a dependent variable. For instance, an experiment might explore the extent to which the writing quality of final drafts of student papers is affected by the kind of instruction they received. In this case, the dependent variable would be writing quality of final drafts.

Independent Variables

In an experimental study, a variable that determines (or causes) the score of a dependent variable is called an independent variable. For instance, an experiment might explore the extent to which the writing quality of final drafts of student papers is affected by the kind of instruction they received. In this case, the independent variable would be the kind of instruction students received.

Probability

Beginning researchers most often use the word *probability* to express a subjective judgment about the likelihood, or degree of certainty, that a particular event will occur. People say such things as: "It will probably rain tomorrow." "It is unlikely that we will win the ball game." It is possible to assign a number to the event being predicted, a number between 0 and 1, which represents *degree of confidence* that the event will occur. For example, a student might say that the likelihood an instructor will give an exam next week is about 90 percent, or .9. Where 100 percent, or 1.00, represents certainty, .9 would mean the student is almost certain the instructor will give an exam. If the student assigned the number .6, the likelihood of an exam would be just slightly greater than the likelihood of no exam. A rating of 0 would indicate complete certainty that *no* exam would be given(Shoeninger, 1971).

The probability of a particular outcome or set of outcomes is called a *p-value* . In our discussion, a p-value will be symbolized by a *p* followed by parentheses enclosing a symbol of the outcome or set of outcomes. For example, p(X) should be read, "the probability of a given X score" (Shoeninger). Thus p(exam) should be read, "the probability an instructor will give an exam next week."

Population

A *population* is a group which is studied. In educational research, the population is usually a group of people. Researchers seldom are able to study every member of a population. Usually, they instead study a representative sample – or subset – of a population. Researchers then generalize their findings about the sample to the population as a whole.

Sampling

Sampling is performed so that a population under study can be reduced to a manageable size. This can be accomplished via random sampling, discussed below, or via matching.

Random sampling is a procedure used by researchers in which all samples of a particular size have an equal chance to be chosen for an observation, experiment, etc (Runyon and Haber, 1976). There is no predetermination as to which members are chosen for the sample. This type of sampling is done in order to minimize scientific biases and offers the greatest likelihood that a sample will indeed be representative of the larger population. The aim here is to make the sample as representative of the population as possible. Note that the closer a sample distribution approximates the population distribution, the more generalizable the results of the sample study are to the population. Notions of probability apply here. Random sampling provides the greatest probability that the distribution of scores in a sample will closely approximate the distribution of scores in the overall population.

Matching

Matching is a method used by researchers to gain accurate and precise results of a study so that they may be applicable to a larger population. After a population has been examined and a sample has been chosen, a researcher must then consider variables, or extrinsic factors, that might affect the study. Matching methods apply when researchers are aware of extrinsic variables before conducting a study. Two methods used to match groups are:

Precision Matching

In *precision matching*, there is an experimental group that is matched with a control group. Both groups, in essence, have the same characteristics. Thus, the proposed causal relationship/model being examined allows for the probabilistic assumption that the result is generalizable.

Frequency Distribution

Frequency distribution is more manageable and efficient than precision matching. Instead of one-to-one matching that must be administered in precision matching, frequency distribution allows the comparison of an experimental and control group through relevant variables. If three Communications majors and four English majors are chosen for the control group, then an equal proportion of three Communications major and four

English majors should be allotted to the experiment group. Of course, beyond their majors, the characteristics of the matched sets of participants may in fact be vastly different.

Although, in theory, matching tends to produce valid conclusions, a rather obvious difficulty arises in finding subjects which are compatible. Researchers may even believe that experimental and control groups are identical when, in fact, a number of variables have been overlooked. For these reasons, researchers tend to reject matching methods in favor of random sampling.

Methods

Statistics can be used to analyze individual variables, relationships among variables, and differences between groups. In this section, we explore a range of statistical methods for conducting these analyses.

Statistics can be used to analyze individual variables, relationships among variables, and differences between groups.

Analyzing Individual Variables

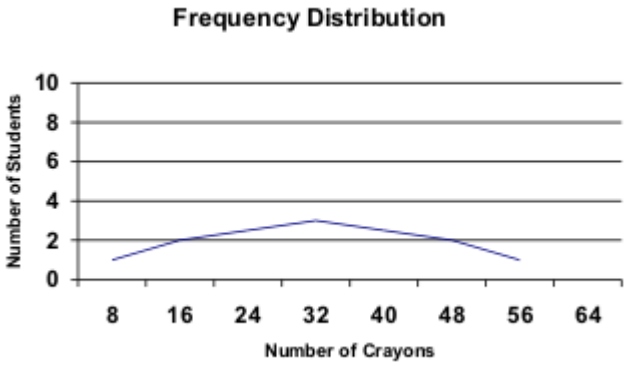
The statistical procedures used to analyze a single variable describing a group (such as a population or representative sample) involve measures of *central tendency* and measures of *variation*. To explore these measures, a researcher first needs to consider the *distribution*, or range of values of a particular variable in a population or sample. *Normal distribution* occurs if the distribution of a population is completely normal. When graphed, this type of distribution will look like a bell curve; it is symmetrical and most of the scores cluster toward the middle. *Skewed Distribution* simply means the distribution of a population is not normal. The scores might cluster toward the right or the left side of the curve, for instance. Or there might be two or more clusters of scores, so that the distribution looks like a series of hills.

Once *frequency distributions* have been determined, researchers can calculate measures of central tendency and measures of variation. Measures of central tendency indicate averages of the distribution, and measures of variation indicate the spread, or range, of the distribution (Hinkle, Wiersma and Jurs 1988).

Measures of Central Tendency

Central tendency is measured in three ways: *mean*, *median* and *mode*. The mean is simply the average score of a distribution. The median is the center, or middle score within a distribution. The mode is the most frequent score within a distribution. In a normal distribution, the mean, median and mode are identical.

Student	# of Crayons
A	8
B	16
C	16
D	32
E	32
F	32
G	48
H	48
J	56



Measures of Variation

Measures of variation determine the range of the distribution, relative to the measures of central tendency. Where the measures of central tendency are specific data points, measures of variation are lengths between various points within the distribution. Variation is measured in terms of range, mean deviation, variance, and standard deviation (Hinkle, Wiersma and Jurs 1988).

The *range* is the distance between the lowest data point and the highest data point. Deviation scores are the distances between each data point and the mean.

Mean deviation is the average of the absolute values of the deviation scores; that is, mean deviation is the average distance between the mean and the data points. Closely related to the measure of mean deviation is the measure of *variance*.

Variance also indicates a relationship between the mean of a distribution and the data points; it is determined by averaging the sum of the squared deviations. Squaring the differences instead of taking the absolute values allows for greater flexibility in calculating further algebraic manipulations of the data. Another measure of variation is the *standard deviation*.

Standard deviation is the square root of the variance. This calculation is useful because it allows for the same flexibility as variance regarding further calculations and yet also expresses variation in the same units as the original measurements (Hinkle, Wiersma and Jurs 1988).

Analyzing Differences Between Groups

Statistical tests can be used to analyze differences in the scores of two or more groups. The following statistical tests are commonly used to analyze differences between groups:

T-Test

A t-test is used to determine if the scores of two groups differ on a single variable. A t-test is designed to test for the differences in mean scores. For instance, you could use a t-test to determine whether writing ability differs among students in two classrooms.

Note: A t-test is appropriate only when looking at *paired* data. It is useful in analyzing scores of two groups of participants on a particular variable or in analyzing scores of a single group of participants on two variables.

Matched Pairs T-Test

This type of t-test could be used to determine if the scores of the same participants in a study differ under different conditions. For instance, this sort of t-test could be used to determine if people write better essays *after* taking a writing class than they did *before* taking the writing class.

Note: A t-test is appropriate only when looking at *paired* data. It is useful in analyzing scores of two groups of participants on a particular variable or in analyzing scores of a single group of participants on two variables.

Analysis of Variance (ANOVA)

The ANOVA (analysis of variance) is a statistical test which makes a single, overall decision as to whether a significant difference is present among three or more sample means (Levin 484). An ANOVA is similar to a t-test. However, the ANOVA can also test multiple groups to see if they differ on one or more variables. The ANOVA can be used to test between-groups and within-groups differences. There are two types of ANOVAs:

One-Way ANOVA: This tests a group or groups to determine if there are differences on a *single* set of scores. For instance, a one-way ANOVA could determine whether freshmen, sophomores, juniors, and seniors differed in their reading ability.

Multiple ANOVA (MANOVA): This tests a group or groups to determine if there are differences on *two or more* variables. For instance, a MANOVA could determine whether freshmen, sophomores, juniors, and seniors differed in reading ability and whether those differences were reflected by gender. In this case, a researcher could determine (1) whether reading ability differed across class levels, (2) whether reading ability differed across gender, and (3) whether there was an interaction between class level and gender.

Analyzing Relationships Among Variables

Statistical relationships between variables rely on notions of correlation and regression. These two concepts aim to describe the ways in which variables relate to one another:

Correlation

Correlation tests are used to determine how strongly the scores of two variables are associated or correlated with each other. A researcher might want to know, for instance, whether a correlation exists between students' writing placement examination scores and their scores on a standardized test such as the ACT or SAT. Correlation is measured using values between +1.0 and -1.0. Correlations close to 0 indicate little or no relationship between two variables, while correlations close to +1.0 (or -1.0) indicate strong positive (or negative) relationships (Hayes et al. 554).

Correlation denotes positive or negative association between variables in a study. Two variables are *positively associated* when larger values of one tend to be accompanied by larger values of the other. The variables are *negatively associated* when larger values of one tend to be accompanied by smaller values of the other (Moore 208).

An example of a strong positive correlation would be the correlation between age and job experience. Typically, the longer people are alive, the more job experience they might have.

An example of a strong negative relationship might occur between the strength of people's party affiliations and their willingness to vote for a candidate from different parties. In many elections, Democrats are unlikely to vote for Republicans, and vice versa.

Regression

Regression analysis attempts to determine the best "fit" between two or more variables. The independent variable in a regression analysis is a continuous variable, and thus allows you to determine how one or more independent variables predict the values of a dependent variable.

Simple Linear Regression is the simplest form of regression. Like a correlation, it determines the extent to which one independent variables predicts a dependent variable. You can think of a simple linear regression as a correlation line. Regression analysis provides you with more information than correlation does, however. It tells you how well the line "fits" the data. That is, it tells you how closely the line comes to all of your data points. The line in the figure indicates the regression line drawn to find the best fit among a set of data points. Each dot represents a person and the axes indicate the amount of job experience and the age of that person. The dotted lines indicate the distance from the regression line. A smaller total distance indicates a better fit. Some of the information provided in a regression analysis, as a result, indicates the slope of the regression line, the R value (or correlation), and the strength of the fit (an indication of the extent to which the line can account for variations among the data points).

Multiple Linear Regression allows one to determine how well multiple independent variables predict the value of a dependent variable. A researcher might examine, for instance, how well age and experience predict a person's salary. The interesting thing here is that one would no longer be dealing with a regression "line." Instead, since the study deals with three dimensions (age, experience, and salary), it would be dealing with a plane, that is, with a two-dimensional figure. If a fourth variable was added to the equations, one would be dealing with a three-dimensional figure, and so on.

Misuses of Statistics

Statistics consists of tests used to analyze data. These tests provide an analytic framework within which researchers can pursue their research questions. This framework provides one way of working with observable information. Like other analytic frameworks, statistical tests can be misused, resulting in potential misinterpretation and misrepresentation. Researchers decide which research questions to ask, which groups to study, how those groups should be divided, which variables to focus upon, and how best to

categorize and measure such variables. The point is that researchers retain the ability to manipulate any study even as they decide what to study and how to study it.

Potential Misuses:

- Manipulating scale to change the appearance of the distribution of data
- Eliminating high/low scores for more coherent presentation
- Inappropriately focusing on certain variables to the exclusion of other variables
- Presenting correlation as causation

Measures Against Potential Misuses:

- Testing for reliability and validity
- Testing for statistical significance
- Critically reading statistics

Annotated Bibliography

Dear, K. (1997, August 28). *SurfStat australia*. Available: http://surfstat.newcastle.edu.au/surfstat/main/surfstat-main.html

A comprehensive site contain an online textbook, links together statistics sites, exercises, and a hotlist for Java applets.

de Leeuw, J. (1997, May 13*). Statistics: The study of stability in variation*. Available: http://www.stat.ucla.edu/textbook/ [1997, December 8].

An online textbook providing discussions specifically regarding variability.

Ewen, R.B. (1988). *The workbook for introductory statistics for the behavioral sciences.* Orlando, FL: Harcourt Brace Jovanovich.

A workbook providing sample problems typical of the statistical applications in social sciences.

Glass, G. (1996, August 26). COE 502: *Introduction to quantitative methods*. Available: http://seamonkey.ed.asu.edu/~gene/502/home.html

Outline of a basic statistics course in the college of education at Arizona State University, including a list of statistic resources on the Internet and access to online programs using forms and PERL to analyze data.

Hartwig, F., Dearing, B.E. (1979). *Exploratory data analysis*. Newberry Park, CA: Sage Publications, Inc.

Hayes, J. R., Young, R.E., Matchett, M.L., McCaffrey, M., Cochran, C., and Hajduk, T., eds. (1992). *Reading empirical research studies: The rhetoric of research*. Hillsdale, NJ: Lawrence Erlbaum Associates.

A text focusing on the language of research. Topics vary from "Communicating with Low-Literate Adults" to "Reporting on Journalists."

Hinkle, Dennis E., Wiersma, W. and Jurs, S.G. (1988). *Applied statistics for the behavioral sciences*. Boston: Houghton.

This is an introductory text book on statistics. Each of 22 chapters includes a summary, sample exercises and highlighted main points. The book also includes an index by subject.

Kleinbaum, David G., Kupper, L.L. and Muller K.E. *Applied regression analysis and other multivariable methods 2nd ed*. Boston: PWS-KENT Publishing Company.

An introductory text with emphasis on statistical analyses. Chapters contain exercises.

Kolstoe, R.H. (1969). *Introduction to statistics for the behavioral sciences*. Homewood, ILL: Dorsey.

Though more than 25-years-old, this textbook uses concise chapters to explain many essential statistical concepts. Information is organized in a simple and straightforward manner.

Levin, J., and James, A.F. (1991). *Elementary statistics in social research, 5th ed*. New York: HarperCollins.

This textbook presents statistics in three major sections: Description, From Description to Decision Making and Decision Making. The first chapter underlies reasons for using statistics in social research. Subsequent chapters detail the process of conducting and presenting statistics.

Liebetrau, A.M. (1983). *Measures of association*. Newberry Park, CA: Sage Publications, Inc.

Mendenhall, W.(1975). *Introduction to probability and statistics, 4th ed.* North Scltuate, MA: Duxbury Press.

An introductory textbook. A good overview of statistics. Includes clear definitions and exercises.

Moore, David S. (1979). *Statistics: Concepts and controversies, 2nd ed*. New York: W. H. Freeman and Company.

Introductory text. Basic overview of statistical concepts. Includes discussions of concrete applications such as opinion polls and Consumer Price Index.

Mosier, C.T. (1997). MG284 Statistics I - notes. Available: http://phoenix.som.clarkson.edu/~cmosier/statistics/main/outline/index.html

Explanations of fundamental statistical concepts.

Newton, H.J., Carrol, J.H., Wang, N., & Whiting, D.(1996, Fall). Statistics 30X class notes. Available: http://stat.tamu.edu/stat30x/trydouble2.html [1997, December 10].

This site contains a hyperlinked list of very comprehensive course notes from and introductory statistics class. A large variety of statistical concepts are covered.

Runyon, R.P., and Haber, A. (1976). *Fundamentals of behavioral statistics, 3rd ed*. Reading, MA: Addison-Wesley Publishing Company.

This is a textbook that divides statistics into categories of descriptive statistics and inferential statistics. It presents statistical procedures primarily through examples. This book includes sectional reviews, reviews of basic mathematics and also a glossary of symbols common to statistics.

Schoeninger, D.W. and Insko, C.A. (1971). *Introductory statistics for the behavioral sciences*. Boston: Allyn and Bacon, Inc.

An introductory text including discussions of correlation, probability, distribution, and variance. Includes statistical tables in the appendices.

Stevens, J. (1986). *Applied multivariate statistics for the social sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Stockberger, D. W. (1996). *Introductory statistics: Concepts, models and applications*. Available: http://www.psychstat.smsu.edu/ **[1997, December 8].**

Describes various statistical analyses. Includes statistical tables in the appendix.

Local Resources

If you are a member of the Colorado State University community and seek more in-depth help with analyzing data from your research (e.g., from an undergraduate or graduate research project), please contact CSU's Graybill Statistical Laboratory for statistical consulting assistance at http://www.stat.colostate.edu/statlab.html.

Citation Information

Shawna Jackson, Karen Marcus, Cara McDonald, Timothy Wehner, and Mike Palmquist.. (1994 - 2012). Statistics: An Introduction. Writing@CSU. Colorado State University. Available at https://writing.colostate.edu/guides/guide.cfm?guideid=67.

Copyright Information