**WHAT IT IS**

Return to Table of Contents

Inferential statistics deal with drawing conclusions and, in some cases, making predictions about the properties of a population based on information obtained from a sample. While descriptive statistics provide information about the central tendency, dispersion, skew, and kurtosis of data, inferential statistics allow making broader statements about the relationships between data.

**WHEN TO USE IT**

Inferential statistics are frequently used to answer cause-and-effect questions and make predictions. They are also used to investigate differences between and among groups. However, one must understand that inferential statistics by themselves do not prove causality. Such proof is always a function of a given theory, and it is vital that such theory be clearly stated prior to using inferential statistics. Otherwise, their use is little more than a fishing expedition.

For example, suppose that statistical methods suggest that on average, men are paid significantly more than women for full-time work. Several competing explanations may exist for this discrepancy. Inferential statistics can provide evidence to prove one theory more accurate than the other. However, any ultimate conclusions about actual causality must come from a theory supported by both the data and sound logic.

**HOW TO PREPARE IT**

The following briefly introduces some common techniques of inferential statistics and is intended as a guide for determining when certain techniques may be appropriate. The techniques used generally depend on the kinds of variables involved, i.e. nominal, ordinal, or interval. For further information on and/or assistance with a given technique, refer to the books and in-house support listed in the Resources section at the end of this module.

- **Chi-square ($\chi^2$) tests** are used to identify differences between groups when all variables nominal, e.g., gender, ethnicity, salary group, political party affiliation, and so forth. Such tests are normally used with contingency tables which group observations based on common characteristics.

  For example, suppose one wants to determine if political party affiliation differs among ethnic groups. A contingency table which divides the sample into political parties and ethnic groups could be produced. A $\chi^2$ test would tell if the ethnic distribution of the sample indicates differences in party affiliation.

  As a rule, each cell in a contingency table should have a least five observations. In those cases where this is not possible, the Fisher's Exact Test should replace the $\chi^2$ test. Data analysis software will usually warn the user when a cell contains fewer than five observations.

- **Analysis of variance (ANOVA)** permits comparison of two or more populations when interval variables are used. ANOVA does this by comparing the dispersion of samples in order to make inferences about their means. ANOVA seeks to answer two basic questions:

— Are the means of variables of interest different in different populations?
— Are the differences in the mean values statistically significant?

For example, during the Welfare Reform audit, SAO staff wanted to test whether the incomes of job training program participants and nonparticipants were significantly different. ANOVA was used to do this.

- **Analysis of covariance (ACOVA)** examines whether or not interval variables move together in ways that are independent of their mean values. Ideally, variables should move independently of one another, regardless of their means. Unfortunately, in the real world, groups of observations usually differ on a number of dimensions, making simple analyses of variance tests problematic since differences in other characteristics could cause observed differences in the values of the variables of interest.

  For example, suppose auditors/evaluators are comparing the income of job program participants and non-participants. Assuming job program participants were found to have higher incomes, could one conclude the explanation was their participation in the program? Although this may be the case, one cannot yet draw that conclusion. Other distinguishing characteristics of program participants might explain higher income. In order to control for the effects of those other variables, an analysis of covariance is necessary.

  During the Welfare Reform audit noted above, SAO staff found that the incomes of job training program participants and non-participants differed by gender and by target group classification (federal classification for welfare intervention purposes). An ACOVA was done to determine if the wages of program participants and nonparticipants were still significantly different, after controlling for differences in gender and target group classification. Thus, here gender and target group were considered the covariates.

- **Correlation ($\rho$)**, like ACOVA, is used to measure the similarity in the changes of values of interval variables but is not influenced by the units of measure. Another advantage of correlation is that it is always bounded by the interval:

  $$-1 \leq \rho \leq 1$$

  Here -1 indicates a perfect inverse linear relationship, i.e. y increases uniformly as x decreases, and 1 indicates a perfect direct linear relationship, i.e. x and y move uniformly together. A value of 0 indicates no relationship. Note that correlation can determine that a

relationship exists between variables but says nothing about the cause or directional effect. For example, a known correlation exists between muggings and ice cream sales. However, one does not cause the other. Rather, a third variable, the warm weather which puts more people on the street both to mug and buy ice cream is a more direct cause of the correlation.

As a rule, it is wise to examine the correlations between all variables in a data set. This both warns auditors/evaluators about possible covariations and suggests areas for possible follow-up investigation.

- **Regression analysis** is often used to determine the effect of independent variables on a dependent variable. Regression measures the relative impact of each independent variable and is useful in forecasting. It is used most appropriately when both the independent and dependent variables are interval, though some social scientists also use regression on ordinal data. Like correlation, regression analysis assumes that the relationship between variables is linear.

  Regression analysis permits including multiple independent variables. For example, during the Welfare Reform audit, SAO staff wanted to predict the percentage of JOBS program clients in a given county who had gotten a job. The poverty rate and population density of the county were used as independent variables. These two variables enabled the auditors to predict almost perfectly the percentage of people who got a job.

- **Logistic regression analysis** is used to examine relationships between variables when the dependent variable is nominal, even though independent variables are nominal, ordinal, interval, or some mixture thereof. Suppose that one wanted to determine which program interventions were associated with a JOBS Program client's ability to get a job within six months of exiting the program. The outcome variable would be "job" or "no job," clearly a nominal variable. One could then use several independent variables such as GED completion, job training, post-secondary education and the like to predict the odds of getting a job. Such a method was applied to the JOBS Program audit.

- **Discriminant analysis** is similar to logistic regression in that the outcome variable is categorical. However, here the independent variables must be interval. In the audit of the Probation System, SAO staff explored how well a probationer's rating on drug abuse severity, social adjustment, and similar characteristics predicted whether or not the probationer committed another crime using continuous ratings. The predictive power of these ratings was just slightly better than chance. Of special interest was the percentage of times the model predicted the

probationer would not commit another crime when he or she actually did. Since that percentage was quite high, the model was a poor one.

- **Factor analysis** simultaneously examines multiple variables to determine if they reflect larger underlying dimensions. Factor analysis is commonly used when analyzing data from multi-question surveys to reduce the numerous questions to a smaller set of more global issues.

  For example, during the Department of Information Resources (DIR) audit, SAO staff administered a survey to agencies that used DIR services. The survey had approximately 30 questions, each of which contained three sub-parts, for a total of 90 objective questions. Reporting the responses to all 90 questions was time-consuming and risked losing continuity. Through factor analysis, the audit team found that the survey could be reduced to five central underlying questions. Thus, in many ways, factor analysis is analogous to the process of "rolling up" audit issues.

- **Forecasting** exists in many variations. The predictive power of regression analysis can be an effective forecasting tool, but time series forecasting is more common when time is a significant independent variable.

  Time series forecasting is based on four components:
  - — trends, as indicated by a relatively smooth pattern in the data over a long period of time, e.g., population increase
  - — cyclical effects, as indicated by a wave-like pattern in the data moving above and below a trend line of over one year in duration, e.g., recessions
  - — seasonal effects, as indicated by changes similar to cyclical effects but over a period of less than one year, e.g., expenditures for sand to apply to roads in icy conditions
  - — random variations, as indicated by any change in the data not attributable to any other component.

Time series forecasting models can be expressed as an additive or multiplicative function of these four components. Once measurements of the four components exist, they can be fit to a regression line to predict future values.

**ADVANTAGES**

Inferential statistics can:
- provide more detailed information than descriptive statistics
- yield insight into relationships between variables
- reveal causes and effects and make predictions
- generate convincing support for a given theory
- be generally accepted due to widespread use in business and academia

**DISADVANTAGES**

Inferential statistics can:

- be quite difficult to learn and use properly
- be vulnerable to misuse and abuse
- depend more on sound theory than on implications of a data set