CHAPTER 7

RESEARCH DESIGN

7.1 INTRODUCTION

The choice of a research strategy, demarcation of a population, specific sampling procedure and the use of appropriate statistical methods for data analysis are of utmost importance in the research process. Suitable and proper research design, sampling methods, and statistics ensure a soundly based structured and systematic approach to scientific knowledge which can be checked for accuracy and the ability to generalize results to the population as a whole. In this chapter the choice of an appropriate research design is discussed. The population is demarcated and thereafter attention is given to the sampling and determination of a proper sample size which will be representative of the population under investigation enabling the researcher to generalize the findings. Lastly the statistical methods for the analysis of data as well as the computer programmes that will be used, are reviewed.

7.2 THE RESEARCH APPROACH

The aim of this study, as stated in Chapter 2, is a comprehensive analysis of the work values expressed by the labour force in two sectors of the mining industry. The extent of the internal/external locus of control as determinants of human behaviour in organizational and industrial settings and the investigation of specific differences in the orientation of control are two further objectives.

These differences will be highlighted in terms of ethnicity, language, religion, educational level, income, years of formal schooling received, ocupational level, age, country of origin and economic sector employed in as independent variables. The four value dimensions and the locus of control will be used as the dependent variables and the biographic and demographic variables will be considered as possible nuisance variables. The achievement of the aims of the study entails obtaining information directly from the work force about their work values and their behaviour. Information thus has to be obtained from the subjects by posing questions about their preferences, feelings and behavioural intentions in regard to work values and completed actions that would reveal information about their internal/external locus of control. The resulting information can and may be obtained by way of questionnaires. Because of the location of the sample and the availability of subjects at training venues, the survey method of data gathering is considered as the appropriate method.

Although the basic approach for this research is the survey method, the data will be considered as being part of an experiment with multiple factorial analysis of variance as the main statistical method. In this instance the research approach would be an a posteriori quasi-experimental design involving questionnaires.

7.2.1 A POSTERIORI QUASI-EXPERIMENTAL DESIGN

A posteriori or post hoc comparison refers to a comparison of means which has not been pre-planned but which allows the researcher to analyse the data in order to ascertain the differences attributable to various independent variables which have given rise to significant F-ratios[1] (Shavelson, 1981, p 469). A posteriori comparison may be defined as a "hypothesis testing of the differences among population means carried out following an analysis of variance" (Bohrnstedt and Knoke, 1988, p 236). The basic requirement for using post hoc comparisons is that the overall F in the analysis of variance must be significant. Post hoc comparisons enable the researcher to make multiple comparisons among a set of means by utilizing the notion of a contrast.[2] The best known a posteriori multiple means comparison test is the Scheffé test (Bohrnstedt et al, 1988, pp 236-238).

A quasi-experimental design is a research plan that has some but not all the validity features of an experimental design. The researcher does not necessarily assign subjects randomly to treatment and control conditions and manipulations of the independent variable are quite difficult if not impossible under certain circumstances (Dooley, 1990, pp 182, 198). There is a change of emphasis in quasi-experimental designs so that the issue is whether an independent variable is an indicator of whatever the real cause may be and not the actual

[1] F-ratio is a test statistic formed by the ratio of two mean-square estimates of the population error variance.

[2] A contrast is a set of weighted population means that sum to zero used in making post hoc comparisons of treatment groups. It is generally labelled ψ (psi).

cause of the dependent variable (Dane, 1990, p 105). Mason et al (1989, p 127) view quasi-experimental design as an alternative to experimental design in that it can be carried out in field settings and does not need to comply with the requirement of equalization of groups by means of the random assignment of subjects.

The researcher may be forced to use this type of design because the study may have to be carried out in the natural setting where the experimental event occurs as is the case in hand. The researcher wishes to determine the effect of the independent variable on the dependent variable(s) and also, if possible, the confounding influence of the nuisance variables. The researcher does not have total control in these natural settings and therefore employs quasi-experimental techniques which were developed to deal systematically with the threats of internal and external validity (Mason et al, 1989, p 127). Internal validity refers to the truthfulness of the claim of a causal linkage between variables internal to the design while external validity points to the extent to which research findings may be applicable to other populations, other times and other settings (Dooley, 1990, pp 183, 218). Quasi-experimental designs are susceptible to the threats of regression, history, maturation, testing and instrumentation.

Instrumentation effects refer to changes in the manner in which the dependent variable is measured. Testing effects point to changes in responses caused by measuring the dependent variable. Maturation involves systematic changes over time regardless of specific events. The history effect refers to the effect produced whenever some uncontrolled event alters participants' responses. Regression refers to the tendency for extreme behaviour to be replaced by less dramatic behaviour (Chadwick, Bahr and Albrecht, 1984, pp 178-179; Dane, 1990, pp 331-339). Dane (1990, pp 106-112) and Mason <u>et al</u> (1989, pp 129-137) advance several quasi-experimental techniques, viz nonequivalent control-group design, counterbalanced design, equivalent-time-samples design, time-series design and regression-discontinuity design. In the nonequivalent control-group design both the experimental and the control group take a pretest as well as a posttest. Only the experimental group is exposed to the experimental variable and is compared to a similar group (control group) not exposed to the experimental variable. This design may be presented as follows (Mason <u>et al</u>, 1989, p 129):

Experimental group O_1 X O_2
Control group O_1 O_2
O_1 = pretest
X = experimental intervention
O_2 = posttest.

The counterbalanced design is one in which there are several treatments and several respondents and each respondent is presented with each treatment condition in random order. The equivalent time-samples design refers to an experimental situation in which each subject serves repeatedly under the experimental and control conditions. The design may involve the alternation of experimental and control conditions. The time-series design is an extended repeated measures design in which the dependent variable is measured several times before and after the introduction of the independent variable. Time-series design, of which there are two types, viz interrupted time-series and multiple time-series, implies a series of measures taken over a period of time (Dane, 1990, p 106; Howard, 1985, pp 123-126). Regression-discontinuity design is a cross-sectional design involving one measurement of different groups that represent different time periods. It is especially appropriate when people or groups are given awards or those in need given extra help and the researcher would like to discover the consequences of such interventions (Howard, 1985, p 121).

The case in hand may be termed a one shot case study and may be represented by the following formula:

X ----- O
X = exposure to the experimental variable
O = observation of the group (measurement).

Only one measurement is taken to determine the influence of the main effect and the interaction between the main and side effects. There is no control group but the independent variable(s) has a controlling effect.

7.2.2 SURVEY RESEARCH

Dane (1990, p 338) defines survey research as a method of "obtaining information directly from a group of individuals". Chadwick et al (1984, p 442) view it "as a research technique that puts questions to a sample of respondents by means of a questionnaire or an interview". Mason et al (1989, p 52) see survey research "as a technique to study the distribution of characteristics in a population". Random assignment, manipulation of the independent variable and testing of the cause-effect hypothesis seldom form part of survey research. The size of the sample, which in survey research is generally large, distinguishes it from other research strategies and methods (Dane, 1990, p 120). The three main methods of survey research are self-administered questionnaires, interview surveys and telephone surveys (Babbie, 1989, p 238; Baker, 1988, p 168).

The survey research process starts with the selection of valid measurements. A valid measurement is a questionnaire containing questions that measure the concept(s) which the researcher intends. Therefore the questions must be worded carefully and unambiguously. The gap between what the researcher wants to measure and the results of the survey must be as narrow and as small as possible (Baker, 1988, p 166). Having selected the appropriate test or having constructed an appropriate questionnaire to measure the concepts the researcher wants to measure, the researcher decides upon modes of eliciting information from the respondents. In the case in hand, questionnaires were chosen to record the respondents' answers. Thereupon respondents were selected. The relevant criterion in selecting respondents is that the population should be appropriate to the questions asked or, put another way, the questions should apply to the population from which the respondents are selected. The questions must be acceptable to the respondents, not give offence and be easily understood by respondents.

The next step in the research process is the administering of the survey. In the case in hand the questionnaires were distributed personally to randomly selected subjects on two mines. The administration was done by two personnel managers. They also gave instructions on how to complete the questionnaire. Once the questionnaires were completed, the respondents had to return them personally to the personnel managers.

7.2.3 THE SURVEY RESEARCH PROCESS

Baker (1988, pp 174-175) discusses four types of questions which may form part of a questionnaire, i.e. closed-ended questions, open-ended questions, contingency questions and matrix questions. Examples of matrix questions are the response categories of a Likert scale. The respondents select a response from a set of five or six response categories. The Value Survey Module and the Activism and Powerful Others-scale of the instrument used in this research (Work Value Survey Module) are both Likert scales. Open-ended questions allow for a more detailed written answer in the space provided. Baker (1988, p 174) suggests that a specific number of lines be left open, for answers to draw a more demarcated and precise reply. Too many lines may lead to the respondent skipping the items. Interesting questions should be put first. It may encourage the respondent to fill out the questionnaire. Sensitive questions should come near the end of the questionnaire (Baker, 1988, p 173). The researcher should take care that the questions are worded in such a way that the respondents understand them. Also, the set of questions should be designed in such a way that they really tap the attitudes towards, and measure the topic concerned (Baker, 1988, p 168).

7.2.4 ADMINISTERING THE QUESTIONNAIRE

Chadwick et al (1984, p 147) advance two broad strategies for collecting data by self-administered questionnaires. Questionnaires may be hand-delivered to individual respondents and collected after a few days or they may be administered to groups. The second strategy, according to Chadwick et al (1984, p 147), is much more efficient. Not only is data collection made easier but it enables the instructor to explain the purpose of the questionnaire and the instructions for their completion, handle enquiries and appeal to and motivate respondents to participate and complete the questionnaire. Although the questionnaire is administered to a group, every respondent completes it privately.

7.2.5 ADMINISTRATION OF THE WORK VALUE SURVEY MODULE

The Work Value Survey Module (the Value Survey Module and the Activism and Powerful Others-scale) was administered to a group of randomly selected employees at each mine. As it was difficult for the personnel manager in question to assemble the group in toto, he visited each respondent, handed out the questionnaire, explained the purpose and the aim of the research as well as the necessary instructions for the completion of the questionnaire. The instructor (personnel manager) also motivated participants in the research. On completion, the respondents handed the questionnaires directly to the instructor.

7.3 POPULATION

De la Rey (1978, p 16) specifies a population as "all the species, persons, or objects being present at a certain place and time holding a specific characteristic". It is clear from this definition that a population includes all possible members. De la Rey calls it a total aggregate. In order that the demands of scientific verification may be satisfied, it is necessary to demarcate and define the population as precisely as possible (De la Rey, 1978, p 16). All the employees working on a gold mine of the Anglo American Corporation in the Western Transvaal and all the employees working on a coal mine of Iscor in the North-Western Transvaal, made up the population for this research. It included employees on all levels ranging from unskilled blue-collar labourers to the managerial level, belonging to different racial and ethnic groups. The educational qualifications vary from primary school education to post-graduate level.

7.4 SAMPLING PROCEDURE

De la Rey (1978, p 16) views a sample as "a smaller number of persons, which one way or another, were drawn from a demarcated population for participation in a research project". There are different sampling methods and the method utilized may have an effect on the research findings. The way the sample is drawn must ensure that the characteristics of the population be present to the same extent in the sample. Such representativeness can be guaranteed only by drawing the sample structurally and methodically enabling the researcher to obtain reliable results (De la Rey, 1978, p 16). In the case in hand the sample consists of 400 employees, 200 from each mine randomly selected in each case from an alphabetical list of employees supplied by the companies.

7.4.1 SAMPLING METHOD

The method was determined to a great extent by the availability of subjects as participants. The researcher attempted to secure a large sample as availability affects the representativeness of the sample negatively. The method of sampling may be described as systematic sampling. According to De la Rey (1978, p 21) if a population or members of a population are classified randomly one way or another in regular succession, a systematic sample may be taken. Systematic sampling is a probability sample selecting every nth person after a random start. On the gold mine a number was randomly selected between 1 and 60, viz 27, and thereafter every 30th name until 200 had been selected. On the coal mine a number was also randomly selected between 1 and 40, viz 25, and henceforth every 20th name until 200 names had been selected. To ascertain the sufficiency and representativeness of the sample, Mc Hugh's formula was employed. McHugh (Guion, 1965, p 126) offered the formula

$$N = 3 + \frac{Z^2}{d^2}$$ where

N = number of subjects required

d = the permissible deviation from the population correlation coefficient and

Z = the normal deviate value of the desired confidence level.

The researcher accepted 5% as the lowest confidence level and equalized it with a value of 1,96. The deviation from the population correlation coefficient was put at 0,10. Ghiselli (in De la Rey, 1976, p 163) proved that the correlation coefficient seldom, if ever, exceeded 0,50. A deviation of 0,10 is therefore acceptable. Substituting the values of 1,96 and 0,10 in Mc Hugh's formula yield a N of 387. A sample of 400 exceeds this number and consequently is acceptable as sample of adequate size and hence it is hoped that the resulting statistical calculations ought to yield reliable and representative results and conclusions.

7.5 STATISTICAL METHODS

Data will be extensively analysed according to criteria developed and expressed by Ferguson (1981), Tabachnick and Fidell (1983), Ott and Mendenhall (1990), Shavelson (1981) and Harris (1975). The major tools of analysis may be descriptive statistics, correlational statistics, analysis of variance, Student's t-test, Kruskal-Wallis non-parametric oneway analysis of variance, Hotelling's T^2-test, discriminant analysis and the Mann-Whitney U-test[3]. The researcher hopes to ascertain the existence of significant differences or not between the cultures of a private enterprise and a semi-state corporation. The researcher also hopes to ascertain the influence of independent or moderator variables such as sex, religion, educational level, years of formal education received, income, occupational level, age, country of origin, economic sector and ethnic affiliation on work values and locus of control.

[3] Factor analysis will not be discussed in detail in chapter 6 as it would not be applied as a statistical tool of analysis of data. It is only applied to examine the structure of the questionnaire with an eye on validity.

7.5.1 ANALYSIS OF VARIANCE

Bohrnstedt <u>et al</u> (1988, p 219) define analysis of variance (Anova) as " a statistical test of the difference of means for two or more groups". Ott <u>et al</u> (1990, p 695) view Anova as "a procedure for comparing more than two population means" while Ferguson (1981, p 234) holds that Anova is "a method for dividing the variation observed in experimental data into different parts, each part assignable to a known source, cause or factor". Anova is thus a method to statistically ascertain whether or not differences between two or more groups exist. The variance is partitioned into variance between groups

$$\sigma^2 = \frac{n\Sigma d^2}{r - 1} \quad . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; A$$

and variance within groups

$$\sigma^2 = \frac{\Sigma\Sigma x_i^2}{r(n-1)} \quad . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; B$$

and is usually expressed as the ratio $\frac{A}{B}$ being called the F-ratio (Du Toit, 1963, p 108). However, it is not only of importance to know, when two or more means are compared, whether there are any reliable differences among them. Especially in experimental research using discrete groups as (the) independent variable(s) the major research question usually focuses on the extent to which the dependent variables differ as a function of group membership. Also, in all techniques for testing group differences, it is possible to test the strength of association between independent and dependent variables. Depending on the number of independent variables and dependent variables, a variety of techniques are available to undertake an analysis of variance. However, the logic behind an analysis of variance may be explained as follows: The Anova model tests the null hypothesis (H_o) that all sample means are drawn from the same population and therefore are equal. The H_o may be represented as $H_o : \mu_1 = \mu_2$ ..... $= \mu_j$. This implies that the group mean will equal the grand mean μ. To measure the effect of an independent variable (IV) on a dependent variable (DV), cognizance has to be taken of this fact. "Effect" may be defined as the impact on a DV of being in a certain treatment group. The Anova model revolves around the question of how much of the total variation in the dependent variable can be explained by the IV or "treatment" variable and how much is left unexplained. The general Anova model with one IV may be presented as

$$Y_{ij} = \mu + a_j + e_{ij}$$

where e_{ij} = error term.[4]

This formula, according to Bohrnstedt <u>et al</u> (1988, p 222), indicates that the score of observation i, which is also a member of group j (hence Y_{ij}), is a function of a group effect, a_j, plus the population mean and random error, e_{ij}. The numerator of the sample variance is then partitioned into two independent additive components to enable the researcher to estimate the proportion of variance in Y_{ij}.

[4] Error term is the difference between the observed score and the score predicted by the model.

The formula $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ is applied to divide the numerator into two components.

$\sum_{i=1}^{N} (Y_i - \bar{Y})^2 = \sum_{j=1}^{J} \sum_{i=1}^{nj} (Y_{ij} - \bar{Y})^2$ as the sum of observations across the J subgroups or treatments equals the total sample size N. The term

$\sum_{j=1}^{J} \sum_{i=1}^{nj} (Y_{ij} - \bar{Y})^2$ is called the total sum of squares (SS_{total}) and is partitioned into a between sum of squares ($SS_{between}$) and a within sum of squares (SS_{within}). Variance is thus expressed as the F-ratio

$$\frac{MS_{between}}{MS_{within}}$$

The total sum of squares refers to a number obtained by subtracting the scores of a distribution from their mean, squaring and summing these values. Between sum of squares is a value obtained by subtracting the grand mean from each group mean, squaring these differences for all individuals and summing them. Within sum of squares refers to the value obtained by subtracting each subgroup mean from each observed score, squaring and summing them (Bohrnstedt _et al_, 1988, pp 219-224; Ott _et al_, 1990, pp 527-540). Dividing the $SS_{between}$ and SS_{within} by their respective degrees of freedom, provide the $MS_{between}$ and the MS_{within} with which the F-ratios may be calculated.

The different techniques of analysis of variance are one-way analysis of variance, factorial anova, one-way manova and factorial manova. An one-way classification of variance enables the researcher to measure the effect of an independent variable on (a) dependent variable (s) (Ferguson, 1981, p 235). In factorial Anova two independent variables or experimental variables are simultaneously investigated. It involves two bases of classification. These classification variables in analysis of variance, are called factors. Because there are two factors, the design is termed a "two-way design". (There might be three or more factors but the larger the design the more difficult the interpretation of results). The two-way design contains an effect term for each factor and a term for the interaction effect produced by both factors operating simultaneously.

Each score is considered to be influenced by its row, column and cell. Effects due to either column or row are called main effects while the effects due to column and row in combination are called interaction effects (Mason et al, 1989, p 231). Main effects are thus due to a single factor while interaction effects refer to influences of two or more factors in combination.

In a two-way factorial Anova the total sum of squares is partitioned into three parts, viz a between-rows sum of squares, a between-columns sum of squares and an interaction sum of squares. The total sum of squares of all observations about the grand mean is

$\sum_{r=1}^{R} \sum_{c=1}^{c} \sum_{i=1}^{n} (X_{rci} - \overline{X}....)^2$ (Ferguson, 1981, p 253).

However, with more than one measurement for the treatment combinations (experimental conditions), the total sum of squares may be divided into four additive components, viz a between-rows sum of squares, a between-columns sum of squares, an interaction sum of squares and a within-cells sum of squares. The variance is expressed as the ratio of the interaction effects (S_{rc}^2) to the within-cells effect (S_w^2), viz

$$F_{rc} = \frac{S_{rc}^2}{S_w^2}$$

(Ferguson, 1981, pp 252-266).

Multivariate analysis of variance (one-way Manova) is designed to investigate differences among two or more levels of an independent variable in terms of their impact on a set of dependent variables. Factorial Manova implies the extension of Manova to research comprizing more than one independent variable (Tabachnick et al, 1983, p 58). Manova has the advantage that the measuring of several dependent variables may improve the chance of discovering changes produced by different treatments and interactions. Manova may also reveal differences not shown in separate Anovas. However, the analysis is quite complex. In factorial Manova, a " 'best linear combination' of dependent variables is formed for each main effect and interaction. The combination of dependent variables that best separates the groups of the first main effect may be different from the combination that best separates the groups of the second main effect or the cells from an interaction" (Tabachnick et al, 1983, pp 222-223).

Manova is also subjected to the limitations of unequal sample sizes, multivariate normality, outliers, linearity, multi-collinearity and singularity and homogeneity of variance - covariance. These limitations are discussed in detail under the heading "Discriminant analysis".

Manova revolves around research questions such as: Are changes in behaviour associated with different levels of an independent variable due to something other than random fluctuations or individual differences occurring by chance (main effects of independent variables) and do independent variables interact in their effect on behaviour (interactions among independent variables)? (Tabachnick et al, 1983, pp 226-227). According to Tabachnick et al (1983, pp 235-238) an appropriate data set for Manova should contain one or more independent variable(s) (classification variables) and two or more dependent variables (measures) on each subject or sampling unit within each combination of independent variables. Each independent variable may have two or more levels. The Manova equation for equal n can be developed through extension from Anova. Anova involves the partitioning of the total variance into two independent additive components, viz sum of squares between groups and sum of squares within groups. For factorial designs the variance between groups can be further partitioned into variance associated with the first independent variable, variance associated with the second independent variable and variance associated with the interaction between the two independent variables. Each n is the number of scores composing the relevant marginal or cell mean or $SS_{bg} = SS_D + SS_T + SS_{DT}$ (Tabachnick et al, 1983, p 238).

Analysis of variance may also be used to conduct a profile analysis as Anova is analogous to the parallelism test, levels test and flatness test (discussed under Hotelling's T^2-test). Treatments correspond to rows, the dependent variables to columns and the interaction between columns and rows is also assessed (Harris, 1975, p 81).

Multiple comparison techniques allow the researcher to investigate post hoc hypothesis involving the means of individual groups or sets of groups. Examples of multiple comparison techniques are the Duncan test, the T-test, Tukey's test, Bonferroni test and the Scheffé-test. The Scheffé-test is the most popular and is a relatively conservative multiple comparison technique (Shavelson, 1981, p 470; Howell, 1989, p 240). Multiple comparisons by means of the Scheffé-test may be conducted regardless of whether the overall F is significant. Howell (1989, p 235) presented the formula

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{MS_{error}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

with degrees of freedom (df) equal to the number of groups - 1 and $N_1 + N_2 - 2$ in order to perform the Scheffé-test. The specific approach used towards the calculation of the post hoc Scheffé-test, describing the data, is that of Horvath (1985, p 226). It is similar to the method described by Howell (1989, p 236-240) but differs in terms of the formula by which the critical values in the F-tables are determined. Horvath uses the normal critical F-values while Howell's approach is similar, except that the obtained F-ratio is multiplied by a factor of (k - 1) where k is equal to the number of groups or subgroups (i.e. the row-effect).

7.5.2 HOTELLING'S T^2-TEST

Hotelling's T^2-test enables the researcher to compare two groups on several variables simultaneously (De la Rey, 1978, p 71). As in the case of Student's t-test, Hotelling's T^2 can be employed to test a single group or two independent groups (Harris, 1975, p 67). According to Tabachnick et al (1983, p 56) Hotelling's T^2 is a special case of multivariate analysis of variance (as the t-test is a special case of univariate analysis of variance) in which two groups comprise the independent variable. Hotelling's T^2 is applied to determine whether the groups differ on a set of dependent variables. Hotelling's T^2 determines whether the centroids (combined averages on the dependent variables) differ for the two groups. Harris (1975, p 78) offers the following formula to compute Hotelling's T^2:

$$T^2 = [N_1N_2/(N_1+N_2)](\overline{X}_1-\overline{X}_2)^1S_C{}^{-1}(X_1-X_2)$$

There is no evidence relating to the robustness of T^2 except that large sample sizes are needed. When the dependent measures originated from a normal distribution, the computed T^2-values conform to the F-distribution (Harris, 1975, p 87).

Certain assumptions, however, have to be met before a T^2-analysis of data may be conducted (Harris, 1975, pp 85-88). The averaging together of the covariance matrices for two groups (the independent variable) before conducting a T^2-analysis of the differences between two groups, involves the implicit assumption that the differences between S_1 and S_2 simply represent random fluctuations about a common population covariance matrix Σ. The null hypothesis (H_o) includes both the hypothesis that $\mu_1 = \mu_2$ and that $\Sigma_1 = \Sigma_2$. However, the second hypothesis is only an assumption on which the correctness of the validity of the first one depends. Rejection of the H_o thus could be due to the fact that $\Sigma_1 \neq \Sigma_2$ rather than to non-null differences between μ_1 and μ_2. Hotelling's T^2 is more sensitive to difference in means than to differences in variances and covariances and the true significance level of T^2 is unaffected by discrepancies between Σ_1 and Σ_2 as long as the sample sizes are fairly large and $N_1 = N_2$ (Harris,1975, p 85). The symbol Σ refers to the comman population covariance matrix.

In some situations the entries in the population variance-covariance matrix are a priori specified (preplanned). The observed variances could be uniformly larger than the hypothesized values suggest. The individual differences in choice probability are inflating the response variabilities. The researcher should therefore be careful to apply formulas for the mean and variance of a multi-nominal distribution to situations where the assumption that all S_s have the same generating probability (ties) is unlikely to be met. According to Harris (1975, p 86) the formula for T^2 is easily corrected to known covariance formulas simply by substituting Σ for S or S_c. The significance of the resulting T^2 is then obtained from the chi-square table with p degrees of freedom. Another assumption on which Hotelling's T^2 is based is that the vectors of outcomes of variables are sampled from a multi-variate normal distribution. As already stated, little is known about the robustness of T^2. For fairly large samples however, computed T^2-values conform to the F-distribution, no matter what shape the parent population takes.

Hotelling's T^2-test is a suitable test to apply in profile analysis[5] as the overall T^2-test for two samples "lumps together two sources of differences between the two groups' response vectors (profiles): a difference in the level of the two curves and differences in the shapes of the two curves" (Harris, 1975, p 80). Methods which analyse these two sources of difference, viz level and shape, separately and in addition, provide a simple test of the flatness of the combined or pooled profile for the two groups are known as profile analysis. Three methods are available in profile analysis to test the response vectors, viz a parallelism test, the levels test and the flatness test (Harris, 1975, pp 80-81). The parallelism approach tests the hypothesis that the profiles of the two groups have the same shape that is $\mu_{\text{slope } 1} = \mu_{\text{slope } 2} = 0$.

In this instance the slope of each line segment making up that profile will be the same for each group. The levels approach tests the hypothesis that the profiles for the two groups are at the same mean level, that is $\mu_{w1} - \mu_{w2} = 0$.

[5] Other methods to determine profile simularities are the method of Du Mas, the method of Du Toit, the method of Osgood and Suci and Cattell's method (Smit, 1991, pp 97-104). However, because these methods are not going to be used in the case in hand, they will not be discussed in detail.

This implies that the aggregali mean of the means of the separate variables is identical for the two groups, which means that the difference between two group means on any variable is zero. The flatness test tests the hypothesis that the pooled profile for the two groups combined is perfectly flat. The combined means are all equal to the same value. The flatness test takes advantage of the fact that a flat profile implies that all line-segment slopes are truly zero (Harris, 1975, p 81).

These three tests are analogous to a two-way univariate analysis of variance in which treatments correspond to rows and response measures (dependent variables) correspond to columns. Harris (1975, p 81) puts it quite aptly:

"The levels test corresponds to a test of the row main effect; the flatness test to a test of the column main effect; and the parallelism test to a test of the interaction between rows and columns. Thus in profile analysis, as in two-way analysis of variance, the interaction test takes precedence with a significant departure from parallelism implying that (a) the two groups must be compared separately on each outcome measure and non-significant departures from the equal levels test hypothesis and or the flatness test hypothesis are essentially non-interpretable since the significant interaction between groups and measures implies that both are significant sources of variation". Greater attention is paid to the concept "Profile analysis" in the next section.

7.5.3 DISCRIMINANT ANALYSIS

A profile analysis may also be done by way of a discriminant analysis.[6] Nunnally (1967, p 372) views profile analysis as "a generic term for all methods concerning groupings of persons". Nunnally proceeds by advancing two major classes of problems in profile analysis, viz that in which the group composition or group membership are known in advance of the analysis and those problems where group membership is not known in advance. The purpose of the analysis in the first instance is to distinguish groups from one another on the basis of scores in a data matrix or scores obtained on a battery of tests. In the second instance the basis of the analysis is to assign individuals to groups in terms of their profile scores.

In the case in hand group membership is known in advance and the purpose of the analysis (discriminant) is to distinguish the groups on the basis of scores in the data matrix. According to Nunnally (op cit, 1967, pp 373-374) profiles have three characteristics, viz level, dispersion and shape. The level of the profile is defined by the mean score of the person over the variables in the profile. The dispersion refers to the extent or

[6] Measures of profile analysis such as measures of profile similarity which entail clustering of variables with factor analysis, measuring the relationship with Bravais-Pearson product-moment correlation and Osgood and Suci's (1952) distance measure D will not be discussed in detail here as the researcher plans to utilize either Hotelling's T^2-test or Discriminant analysis for profile analysis.

degree of divergence from the average. The standard deviation of scores for each person may be seen as a measure of the dispersion. The shape refers to the curve, the high and low points thereof. The method used for clustering profiles in the case in hand is discriminant function analysis. Discriminant function analysis is employed when groups are defined a priori and the purpose of the analysis is to distinguish the groups from one another on the basis of scores obtained in a battery of tests or scores in a data matrix (Nunnally, 1967, p 388). Certain assumptions have to be met before discriminant function analysis can be employed. Discriminant analysis assumes that the predictor variable scores are independently and randomly sampled from a population of scores and that the sampling distribution of any linear combination of predictor variables is normally distributed. This assumption is called multi-variate normality. However, discriminant analysis is robust to failures of normality if the violation is caused by skewness rather than by outliers.

Discriminant function analysis is extremely sensitive to multi-variable outliers. Outliers are cases with extreme values on a variable or combination of variables which unduly influences the averages and variability of scores and invalidates the generalizability of the solution to the population. Therefore outliers have to be eliminated or transformed before discriminant analysis can be performed. The discriminant model also assumes a linear relationship among all predictor variables within each group. Violation of this assumption, however, simply leads to reduced power rather than to an increase in Type I error.[7] The discriminant model is also based on the assumption of homogeneity of variance-covariance. If classification is the goal of the analysis this assumption has to be met. If the sample sizes are quite large, discriminant function analysis displays a robustness in respect of violation of the assumption of equal variance-covariance matrices. With unequal and/or small sample sizes, homogeneity of variance-covariance should be assessed.

Scatterplots of the scores on the first two canonical discriminant functions can also be assessed for each group separately. Scatterplots roughly equal in size give evidence of homogeneity of variance-covariance matrices. The discriminant model also assumes that two variables in a matrix should not be perfectly or almost perfectly correlated (multi-collinearity). Neiter should one score be a linear or nearly linear combination of others (singularity). Multi-collinearity and singularity make the inversion of matrices unreliable (Tabachnick <u>et al</u>, 1983, pp 300-301). The discriminant model is set up so that the first discriminant function maximally separates two groups and the second discriminant function, which operates orthogonally to the first, then separates the remaining groups on the basis of information not accounted for by the first discriminant function (Tabachnick <u>et al</u>, 1983, p 295).

[7] A statistical decision error which occurs when a true null hypothesis is rejected. Its probability is 1 - ∝.

According to Tabachnick et al (1983, p 295) the total number of possible discriminant functions is either one fewer than the number of groups or equal to the number of predictor variables. However, Tabachnick et al are adamant that only the first two discriminant functions significantly and reliably discriminate among groups.

The significance of a set of discriminant functions is established by partitioning the variance in the set of predictors into two sources, viz variance which is attributable to differences between groups and variance attributable to differences within groups (Tabachnick et al, 1983, p 302). Tabachnick et al advance as a fundamental formula for testing the significance, the equation

$$\sum_{ij}\sum(Y_{1j}-GM)^2=n\sum_{j}(\overline{Y}_j-GM)^2+\sum_{ij}\sum(Y_{1j}-\overline{Y}_j)^2$$

and use this procedure to form cross-products matrices in the following way:

$S_{total} = S_{bg} + S_{wg}$ (Tabachnick et al, 1983, pp 237, 302)

The total of cross-products matrices is partioned into cross-products matrices with differences between the two groups (S_{bg}) and differences associated with subjects within groups (S_{wg}). A classification equation is developed for each group to classify cases into groups. According to Tabachnick et al (1983, p 306) each case has a classification score for each group. A case is assigned to the group for which it has the highest classification score. Tabachnick et al (1983, p 306) advance a classification equation

$C_j = c_{jo} + c_{j1}Y_1 + c_{j2}Y_2 + \ldots + c_{jp}Y_p$.

A score on the classification function for group $j(C_j)$ is determined by multiplying the raw score on each predictor variable (Y) by its associated classification function coefficient cj. Then these products are summed over all predictor variables and are added to a constant C_{jo} (Tabachnick et al, 1983, p 306).

There are three types of discriminant function analysis, viz direct discriminant function analysis, hierarchical discriminant function analysis and stepwise discriminant function analysis. The direct discriminant function solves equations simultaneously on the basis of all predictor variables. All the predictor variables enter the equations at once and the dependent variables are considered simultaneously. The hierarchical mode evaluates contributions to group discrimination by predictor variables as they enter the equations in some priority order which is determined by the researcher. This enables the researcher to assess the predictive power of each variable. The researcher may thus determine if the classification of cases to groups improves by adding a specific variable (or a set of variables). When prior variables are viewed as co-variates and the added variable as a dependent variable, this can be seen as an analysis of the covariance. Stepwise discriminant function analysis refers to the determination of the order of entry of variables into the discriminating equation by means of available statistical criteria. The researcher has no a priori reason for ordering entry of variables (Tabachnick et al, 1983, pp 309-313). Stepwise analysis is used for the case in hand. As the researcher does not have a priori reason for ordering the entry of variables into the discriminant equations, statistical criteria, which are available with the stepwise function, have to be applied to determine the order of entry.

The maximum number of discriminant functions extracted within a single discriminant analysis is the lesser of either the number of groups minus one, or equal to the number of predictor variables. However, not all the functions may carry important information. It happens quite frequently that the first few discriminant functions account for the major share of discriminating power with no additional information forthcoming from the remaining functions (Tabachnick et al, 1983, p 318).

Discriminant function plots may be used to interpret the discriminant functions. The discriminant functions are presented by way of pairwise plots of group centroids on all significant discriminant functions. These centroids are the means of obtaining the discriminant scores for each group on each dimension. A discriminant function plot is simply a plot of the canonical discriminant functions evaluated at group means (Tabachnick et al, 1983, pp 313, 319).

Discriminant functions may also be interpreted by examining the loadings of predictor variables on them. Loading matrices are basically factor loading matrices. These factor loading matrices contain correlations between predictor variables and each of the discriminant functions (also called canonical variables) which enable the researcher to name and interpret the functions. Mathematically, the loading matrix "is the pooled within group correlation matrix multiplied by the matrix of standardized discriminant function coefficients" (Tabachnick et al, 1983, p 320).

7.5.4 STUDENT'S T-TEST

Like Hotelling's T^2-test, Student's t-test is also an inferential statistic to test for significant differences between two groups. The two groups may be dependent or independent. Student's t-test enables the researcher to decide whether observed differences between two sample means are caused by chance or represent a true difference between populations (Shavelson, 1981, p 419). De la Rey (1978, p 71) states the following assumptions which have to be met before the t-test can be used:

1. The scores in the respective populations must be normally distributed.
2. As the t-test is based on sample means, the two samples must be big and of equal or almost equal size.
3. The measurements must be on interval or ratio level.
4. The scores in the groups must be randomly sampled from their respective populations.

The use of the t-test also imposes a number of requirements on the collection of data:

1. There is one independent variable with two levels (i.e. groups).
2. A subject appears in one and only one of the groups.

3. The levels of the independent variable may differ from one another either qualitively or quantitatively (Shavelson, 1981, p 421).

Applied to test hypotheses, the purpose of the t-test is to decide whether or not to reject the null hypothesis which is a probabilistic decision as it cannot be made with complete certainty. To determine the probability of observing the difference between the sample means of the two groups under the assumption that the null hypothesis (H_o)[8] is true, a significance test to decide whether the observed sample difference in means has a low probability of occurring in the populations, has to be performed. Bohrnstedt _et al_ (1988, pp 204-205) advance the formula for doing this:

$$s^2 = \frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2}$$

where $N_1 + N_2 - 2$ are the degrees of freedom which are associated with s^2. The value of t is calculated by applying the formula

$$t_{(N_1+N_2-2)} = \frac{(\bar{Y}_2 - \bar{Y}_1) - (\mu_2 - \mu_1)}{s_{(\bar{Y}_2 - \bar{Y}_1)}}$$

$$= \frac{\bar{Y}_2 - \bar{Y}_1}{\sqrt{\dfrac{s^2}{N_1} + \dfrac{s^2}{N_2}}}$$

[8] Null hypothesis (H_o) = no difference between the means of two groups.

Shavelson (1981, p 420) however advances a more simplified formula to determine the t-value

$$t_{\bar{X}_1 - \bar{X}_2} \text{ (observed)} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_{\bar{X}_1 - \bar{X}_2}}$$

Student's t-test assumes that the distribution of variables in the populations from which the samples are drawn, is normal. But it also assumes that the variances in the populations from which the samples are drawn are equal ($\sigma_1^2 = \sigma_2^2$). This is known as homogeneity of variance (Ferguson, 1981, pp 179, 245). According to Ferguson (1981, p 245), moderate departures from homogeneity should not have a serious effect on the inferences drawn from the data. Gross departures from homogeneity, however, may lead to serious errors in the results. Ferguson (1981, p 245) recommends that under circumstances of gross departures from homogeneity, a transformation of the variable which may lead to greater uniformity of variance be used or a nonparametric procedure be applied. Ferguson (1981, p 182) also advances a formula when testing the difference between means for independent samples, assuming homogeneity of variance. A single estimate S^2 is used in calculating the t-value:

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{S^2}{N_1} + \frac{S^2}{N_2}}}$$

However, should the two population variances be different ($\sigma_1^2 \neq \sigma_2^2$), two variance estimates are obtained, viz S_1^2 and S_2^2 which are estimates of σ_1^2 and σ_2^2. The difference is divided by the standard error of the difference and t is computed simply by using the separate variance extimate. The resulting ratio is

$$t' = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}}}$$

This ratio (t') is neither normal nor does it approach a t-distribution.

7.5.5 NON-PARAMETRIC STATISTICS

Two non-parametric statistics are considered, viz Kruskal-Wallis one-way analysis of variance and the Mann-Whitney U-test. Applying non-parametric statistics one or more of certain assumptions have to be met (De la Rey, 1978, p 113):

1. The distribution of scores has to be skewed.
2. Measurement must be on nominal or ordinal level.
3. The sample size must be small ($N \leq 30$).
4. Situations where it is impossible to make certain assumptions in regard to the sample.
5. Situations where it is impossible to realize certain research aims because appropriate parametric statistics are not available.

7.5.5.1 KRUSKAL-WALLIS ONE-WAY ANALYSIS OF VARIANCE

Kruskal-Wallis one-way analysis of variance is applied to help to decide if k independent samples from different populations differ significantly. The samples must be more than two independent samples. The decision is also probabilistic as the problem according to Siegel (1956, p 84) is to determine whether differences among samples represent merely chance variations or signify genuine population differences. Siegel (1956, p 184) observes that the Kruskal-Wallis statistic tests the H_o, that the k-samples come from the same population or from identical populations with respect to averages.

In the computation of the Kruskal-Wallis test the observations or scores are all ranked in a single series. Siegel (1956, p 185) supplies the following formula to calculate the Kruskal-Wallis statistic (H) and observes that if the null-hypothesis (H_o) is true, then H is distributed as chi-square with degrees of freedom = k - 1, provided that the sizes of the various k-samples are not too small:

$$H = \frac{12}{N\ (N + 1\ )} \sum_{j=1}^{k} \frac{R_j 2}{n_j} - 3\ (N + 1)$$

where k = number of samples

n_j = number of cases in jth sample

N = Σn_j, the number of cases in all samples combined

R_j = sum of ranks in the j th sample

$\sum_{j=1}^{k}$ = directs one to sum over the k samples.

7.5.5.2 MANN-WHITNEY U-TEST

The Mann-Whitney U-test is a well-known distribution-free test for two independent samples. Although it is a non-parametric test for comparing the central tendency of two independent samples, it may also be applied to normally distributed populations. Instead of computing means as the sample statistic, however, the Mann-Whitney U-test is based on the ranking of sample scores. Ranking is a sophisticated mathematical operation and can be performed at ordinal level data. The Mann-Whitney U-test tests the H_o that the two samples were randomly drawn from identical populations. This test is especially sensitive to population differences in central tendency.

This H_o is broader than the H_o tested by the corresponding t-test which deals with means of the two samples. The H_o tested by the Mann-Whitney U-test is based on the assumption that the two populations have the same shape and dispersion.

The logic of the Mann-Whitney U-test is quite easy to understand. To compute U, the scores from both samples are pooled and ranked from highest to lowest. Tied observations are then assigned the mean of the rank position they would have occupied had there been no ties. The ranks of observations from group 1 are then summed. Thereupon the ranks for the two samples are totalled and compared. The statistic used in this test, viz the U-value is then given by the number of times a score in one group (with n_2 cases) precedes a score in the other group (with n_1 cases) in the ranking.

If the two samples represent populations not significantly different from each other, then the total ranks should be similar in value. Tied scores are assigned to the average of the ranks they would have had if they had not been tied. The formula to compute U is

$$U = N_1 \, N_2 + \frac{N_1 \, (N_1 + 1)}{2} - \Sigma R_1$$

where ΣR_1 = the sum of ranks for sample 1 (Siegel, 1956, p 120).

On determining the value of U, the test of significance has to be conducted. A z-score is obtained with the aid of the formula

$$Z \text{ (obtained)} = \frac{U - \mu_u}{\sigma_u}$$

where U = the sample statistic

μ_u = the mean of the sampling distribution of sample U's

σ_u = the standard deviation of the sampling distribution of sample U's (Siegel, 1956, p 121),

to find the critical region as marked by Z (critical). Based on Z (critical) the researcher makes a decision to reject or to accept the H_o of no difference (Healy, 1990, pp 193-197; Howell, 1989, pp 300-305).

7.5.6 CORRELATIONAL STATISTICS

Ott et al (1990, p 417) define correlation as a "measure of the strength of the relationship between two variables x and y". The value so obtained is called the coefficient of linear correlation, or simply the correlation coefficient. The stronger the correlation, the better x predicts y. The population correlation coefficient r (rho) is computed as

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$$

This is called the Bravais-Pearson product-moment correlation coefficient. Some textbooks present the formula as

$$r_{xy} = \sqrt{R^2_{y.x}}$$

(Bohrnstedt et al, 1988, p 271)

or

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \cdot \Sigma y^2}}$$

(Du Toit, 1963, p 74).

The Bravais-Pearson product-moment correlation may have a positive or negative sign attached to it to indicate the direction of the correlation. The value of r can range between -1,00 for a perfect inverse association to +1,00 for a perfect positive correlation with zero (r = 0) indicating no relationship at all. Bohrnstedt et al (1988, p 271) see the usefulness of the correlation coefficient in its communication of directionality and magnitude of the association. Ott et al (1990, pp 420-422) note several interpretations of the coefficient of correlation:

1. A correlation coefficient equal to 0,5 does not mean that the strength of the relationship between two variables (x and y) is halfway between no correlation and perfect correlation. The more closely x and y are linearly related, the more the variability in the y-values can be explained by variability in the x-values and the closer r^2 will be to 1. If r = 0,50 the independent variable x is accounting for r^2 = 0,25 or 25% of the total variation in the y-values. r^2 is called the coefficient of determination.[9]
2. X and y could be perfectly related in some way other than in a linear manner when r = 0 or a very small value.
3. Correlations are difficult to add up. The sum of coefficients of correlation does not account for the variability of the y-values about their sample mean.

Spearman's correlation coefficient for ranked data (r_s) may also be calculated. This coefficient of correlation is based on ranked data. Ranking entails separate ranking of a number of items on two dimensions. Based on this ranking, the

[9] The coefficient of determination is a proportional reduction in error statistic (a characteristic of some measures of association which allows the calculation of reduction in errors predicting the dependent variable) for linear regression which expresses the amount of variation in the dependent variable explained or accounted for by the independent variable (Bohrnstedt et al, 1988, p 269).

correlation between the two sets of ranks is determined.[10] Howell (1989, p 110) presents the formula for the calculation of Spearman's rho (r_s) as:

$$r_s = 1 - \frac{6\Sigma D^2}{N(N^2 - 1)}$$

7.5.7 DESCRIPTIVE STATISTICS

Mason <u>et al</u> (1989, p 428) define descriptive statistics as statistics used to summarize data. Bohrnstedt <u>et al</u> (1988, pp 66-81) divide descriptive statistics into measures of central tendency and measures of variation.

7.5.7.1 MEASURES OF CENTRAL TENDENCY

The mode, the median and the mean are measures of central tendency. The mode is the value or category in a frequency distribution that has the largest number, or percentage of cases. The median refers to the value or score that exactly divides an ordered frequency distribution into equal halves, viz the outcome is associated with the 50th percentile. The most frequently used measure of central tendency is the mean which is commonly called the average. The mean is the sum of all scores in a distribution divided by the

[10] Ranked data is data for which the observations have been replaced by their numerical ranks from lowest to highest and Spearman's correlation (r_s) is a correlation coefficient based on ranked data.

number of scores, viz the mean is the arithmetic average. In this research the mean is the measure of central tendency which may be applied to interpret the results of t-scores, discriminant analysis and one-way and other approaches to analysis of variance.

7.5.7.2 MEASURES OF VARIATION

Measures of variation calculated and presented are the range, variance, standard error and standard deviation. The range is an indication of the dispersion based on the difference between the largest and the smallest outcomes in a distribution. The variance is a measure of dispersion for continuous variables indicating an average of squared deviations of scores about the mean and the standard deviation is the square root of the variance and is also used to describe a dispersion of a distribution. The usual way of assigning meaning to the standard deviation is in terms of how many scores fall no more than a standard deviation above or below the mean. For a normal distribution exactly two-thirds of observations lie within one standard deviation of the mean. The standard deviation is basically a measure of the average of the deviations of each score from the mean. The standard error of the mean refers to the standard deviation of sample means in a sampling distribution. It provides information about the amount of error likely to be made by inferring the value of the population mean from the sample mean. The greater the variability among sample means, the greater the chance that inferences about the population mean from a single sample mean will be in error (Shavelson, 1981, p 305).

7.5.7.3 FREQUENCY TABLES

Frequency tables comprise of information about the frequencies across values for biographical variables. The percentage and cumulative percentage will be used to describe and summarize the sample.

7.5.7.4 CROSSTABULATION

A frequency distribution is a useful display of the quantitive attributes of continuous variables or the qualitative attributes of discrete variables. But a crosstabulation (joint contingency table) is "a tabular display of the joint frequency distribution of two discrete variables which has r rows and c columns" (Bohrnstedt et al, 1988, p 101). Thus a crosstabulation indicates the joint outcomes of two variables. The cells which comprise the body of any table show these joint outcomes of two variables. Bohrnstedt et al (1988, p 103) view a cell as "an intersection of a row and a column in a crosstabulation of two or more variables". Marginal distributions consisting of row marginals and column marginals are frequency distributions of each of two crosstabulated variables. Row marginals are the row totals and column marginals are the column totals.

7.6 CONCLUSIONS

In this chapter the research design was discussed. Survey research was discussed in detail and related to the aim of this study. Attention was paid to the process of survey research. The method and procedures for administering the questionnaire and data-collection were discussed. The population was demarcated and the sampling methods and procedure were discussed in detail. Attention was also paid to ascertain an appropriate sample size. Descriptive and inferential statistical methods were explained. The various statistical methods consist of descriptive statistics, different approaches to the analysis of variance, profile analysis (discriminant analysis), the Student's t-test, Hotelling's T^2-test, non-parametric inferential statistics, e.g. Kruskal-Wallis one-way analysis of variance, and Mann-Whitney U-test and correlational statistics which entail the parametric Bravais-Pearson product-moment correlation for normally distributed scores and Spearman's rank correlation which is a non-parametric correlation applicable to ranked data.