Understanding Statistics for the Social Sciences with IBM SPSS

Robert Ho

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Understanding Statistics for the Social Sciences with IBM SPSS

Robert Ho

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

SPSS was acquired by IBM in October 2009

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed on acid-free paper

International Standard Book Number-13: 978-1-138-74228-4 (Hardback)
978-1-138-74220-8 (Paperback)

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

1

Introduction to the Scientific Methodology of Research

1.1 Introduction

Some of the common questions many students ask when they begin their research in psychology are, 'Why do we need to do research? After all, what has the conduct of research got to do with the study of human behavior? Can't we simply study human behavior without recourse to understanding the techniques of scientific investigation, without recourse to hypothesis testing', and of course, the mother of all questions, 'without recourse to statistical analysis?' Fair questions! The simple answer is that without objective, empirical, scientific-based verification of social phenomena, science as we know it will simply cease to exist. And without scientific-based empirical research, progress in knowledge, technology, and innovations will stagnate and possibly come to a screeching halt. At the heart of human existence, evolution, growth, and progress are our needs and desire to learn new things, to understand existing phenomena, to be able to explain scientifically why A causes B, and for many social scientists, to be able to understand 'why people behave the way they do'. To be able to do all these things, we need to be well versed in the scientific methodology of research.

1.2 The Scientific Approach versus the Layperson's Approach to Knowledge

A scientific approach to knowledge is different from a layperson's approach to knowledge. A layperson's approach is typically subjective and based on intuition and everyday observations, whereas a scientific approach to knowledge is based on *systematic observation* and *direct experimentation*. For example, a friend of yours tells you that based on his everyday observations he has concluded that men are smarter than women. Surely, as a budding scientist, you have a lot of questions to ask your friend. How many men and women has he met?

Are these men and women representative of their respective populations? And how did he decide how smart they were? It will not be enough for your friend to say to you that he had met lots of men and women and that he is very good at telling how smart people are. You need proof; show me the evidence! The scientific approach to knowledge does not rely on speculation, guesswork, or armchair philosophizing in drawing conclusions about phenomena. It relies on scientific methodology that incorporates (1) the techniques of *random sampling*, (2) choice of *research designs*, and (3) a thorough understanding of *probability theory* and *hypothesis testing*. For all these methodological issues, statistics plays a crucial role in aiding the researcher to draw inferences/conclusions from the data set about some psychological phenomena.

Before examining the role that statistics plays in scientific research, let's review how these three methodological 'pillars' provide the foundation for the scientific investigation of social/psychological phenomena.

1.2.1 Sampling

Why is the technique of sampling important in research? Sampling is important in social science research because it helps a researcher to generalize results obtained from a specific sample to the population of interest. In conducting research, it is often impossible and not practical to investigate the entire population, unless of course the population is small, like the student population on a university campus. Normally, the research that social scientists conduct is based on very large populations – like the population of Bangkok city, Thailand. To test the entire population of Bangkok residents (~10 million!) will be impossible. As such, research is normally conducted on a sample drawn from the population of interest. However, at the end of the day, it will be necessary for the researcher to be able to generalize the results obtained from the sample back to the population, that is to say, *what is true of the sample must also be true of the population*. To be able to do this, it is critical that the sample selected is a *probability (random) sample* in which all members of a group (population or universe) have an equal and independent chance of being selected into the sample.

1.2.2 Research Designs

When conducting quantitative research, the most common research designs employed can be broadly classified into two types: (a) *between-groups design* and (b) *correlational design*. They differ primarily in terms of the aim of the study and the research questions posed.

1.2.3 Between-Groups Design

When a researcher is interested in investigating whether the manipulation of an independent variable (IV) has some effect on the dependent variable

(DV), then the between-groups design is appropriate. The between-groups design can be further classified into the *univariate* approach and the *multivariate* approach.

1.3 The Univariate Approach

The univariate approach is one in which the research design involves only one DV. For example, if you are interested in investigating whether there is a gender (the IV) difference in problem-solving skills (the DV), and you have a single measure of 'problem-solving skills', then the design is univariate. For example, Table 1.1 shows the problem-solving scores obtained for a sample of 5 male and 5 female subjects.

The univariate design can involve more than one IV and such a design is called a *factorial* design. For example, in addition to investigating gender difference on problem-solving skills, you are also interested in investigating age difference. Suppose that the variable of age is classified into two categories – young/old. Now, the research design involves two IVs (gender and age) and one DV (problem-solving skills). The univariate design will allow for the investigation of the *joint* effect of gender and age on problem-solving skills. More specifically, the univariate design will allow for the 2 × 2 factorial combination of the two variables of gender and age, giving rise to the four groups of 'male-young', 'male-old', 'female-young', 'female-old', and their effects on the DV. For example, Table 1.2 shows the problem-solving scores obtained for a sample of 10 male subjects (5 male-young and 5 male-old) and 10 female subjects (5 female-young and 5 female-old). Regardless of the number of IVs included for testing, the univariate design involves only one DV.

1.3.1 The Multivariate Approach

The multivariate approach is one in which the research design involves more than one DV. For example, apart from investigating whether there is gender

TABLE 1.1

Problem-Solving Scores as a Function of Gender

Male		Female	
s1	65	s1	76
s2	72	s2	65
s3	59	s3	84
s4	89	s4	68
s5	72	s5	82

TABLE 1.2

Problem-Solving Scores as a Function of Gender and Age

Male			Female		
Young		Old	Young		Old
s1	65	76	s1	56	65
s2	72	65	s2	68	89
s3	59	84	s3	72	92
s4	89	68	s4	69	73
s5	72	82	s5	70	81

difference in problem-solving skills, the researcher is also interested in finding whether there are gender differences in mathematics skills, English language skills, and overall grade point average (GPA). The design of the study is multivariate in that it involves more than one DV (the four DVs are the scores of problem-solving skills, mathematics skills, English language skills, and overall GPA). The advantage of the multivariate approach is that it takes into account the interrelations among the DVs and analyzes the variables together. For example, Table 1.3 shows the scores obtained for the four DVs of problem-solving skills, mathematics skills, English language skills, and overall GPA as a function of gender (males vs. females).

A major advantage of the multivariate design is that it is often used in '*repeated-measures*' studies. The design is particularly useful in investigating the effectiveness of an intervention strategy in a pre- and post-study. For example, a teacher may be interested in finding out whether a new learning strategy introduced by the school will be effective in improving the students'

TABLE 1.3

Problem-Solving Skills, Mathematics Skills, English Language Skills, and GPA as a Function of Gender

	Problem-Solving Skills	Maths	English	GPA
Male				
s1	65	84	56	3.2
s2	72	79	64	3.6
s3	59	86	61	3.4
s4	89	93	78	3.8
s5	72	89	64	3.7
Female				
s1	76	64	92	3.5
s2	72	72	89	3.6
s3	59	84	91	3.7
s4	89	71	88	3.6
s5	72	79	95	3.8

TABLE 1.4

Pre- and Post-GPA Scores as a Function of a New Learning Strategy

	Pre-Strategy (GPA Scores)	New Learning Strategy (X)	Post-Strategy (GPA Scores)
s1	2.8	X	3.4
s2	3.1	X	3.3
s3	2.6	X	3.7
s4	2.9	X	3.5
s5	3.0	X	3.6

GPA. A simple multivariate repeated-measures design will answer this question. For example, Table 1.4 shows that the post-strategy GPA scores for a sample of five students have increased over their pre-strategy scores after experiencing the new learning strategy.

1.4 Correlational Design

The correlational approach to investigation is concerned with (1) finding out whether a relationship exists between two variables and (2) determining the *magnitude* and *direction* of this relationship. Unlike the between-groups design, the correlational design does not involve any manipulation of variables, such as manipulating the gender variable – two groups (males and females) – and to see whether there is any difference in problem-solving ability between these two groups. Rather, the correlational approach is concerned with determining whether a naturally occurring set of scores is related to another naturally occurring set of scores. For example, a developmental psychologist may be interested in the relationship between age and height in children. The psychologist selects a group of children for study, and for each child, he/she records their age in years and their height in inches. The psychologist can then calculate the correlation coefficient between these two variables. The correlation coefficient is a number between −1 and +1 that measures both the *strength* and *direction* of the linear relationship between the two variables.

1.5 Hypothesis Testing and Probability Theory

The primary function of hypothesis testing is to see whether our prediction/hypothesis about some social/psychological phenomenon is supported or not. As mentioned earlier, at the very heart of the scientific methodology is an experiment, and part and parcel of conducting an experiment or a research project is to test the predictions/hypotheses generated from a particular theory or past research findings or from our literature review. In testing a hypothesis, data must be collected, analyzed, and interpreted as to whether or not the findings support the study's hypothesis.

1.5.1 Probability

We employ probability to help us make decisions regarding the tests of our hypotheses, that is, whether to accept or reject our hypotheses. When making a decision as to whether to retain or reject a hypothesis, we have to decide at what chance probability level we should reject the hypothesis. For example, if the calculated probability of a test result occurring by chance is very low, then in all likelihood we will reject chance as an explanation for the obtained result and conclude that the result is most likely due to the manipulation of the IV. On the other hand, if the calculated chance probability of occurrence is very high, then we will most likely retain chance as an explanation for the obtained result and conclude that the result is most likely *not* due to the manipulation of the IV but to chance variation.

1.5.2 Statistics and Scientific Research

As mentioned earlier, the scientific methodology incorporates the techniques of random sampling, choice of research designs, and a thorough understanding of probability theory and hypothesis testing. A common tool that is employed in these methodologies to assist the researcher in drawing inferences/conclusions about his/her research is *statistics*.

1.6 Definition of Statistics

Statistics is the science of collecting and learning from data. It is a branch of mathematics concerned with the collection, classification, analysis, and interpretation of numerical facts, for drawing inferences on the basis of their quantifiable likelihood (probability). Statistics also allows the researcher to interpret grouped data, too large to be intelligible by ordinary observation. The field of statistics is subdivided into *descriptive* statistics and *inferential* statistics.

1.6.1 Descriptive Statistics

Descriptive statistics is the branch of statistics that involves organizing, displaying, and understanding data. More specifically, descriptive statistics involves the analysis of data that helps describe, show, or summarize data in a meaningful way such that, for example, patterns might emerge from the data. Descriptive statistics does not, however, allow the researcher to make conclusions beyond the data analyzed or to reach conclusions regarding any hypotheses the researcher might have made. It is simply a way to describe data.

In presenting research findings, descriptive statistics is very important because if raw data were presented alone, it would be difficult to visualize what the data were showing, especially if there were a lot of it. Descriptive statistics, therefore, enables the researcher to present the data in a more meaningful way, which allows simpler interpretation of the data. For example, if we had the IQ scores from 100 students, we may be interested in the average IQ of those students. We would also be interested in the distribution or spread of the IQ scores. Descriptive statistics allows us to do this. Typically, there are two general types of statistics that are used to describe data:

Measures of central tendency—these are ways of describing the *central position* of a frequency distribution for a group of data. In this case, the frequency distribution is simply the distribution and pattern of IQ scores scored by the 100 students from the lowest to the highest. We can describe this central position using a number of statistics, including the *mean*, *median*, and *mode*. For example, the mean is simply the average IQ score – the central position of the distribution of the 100 IQ scores.

Measures of spread—these are ways of summarizing a group of data by describing how spread out the scores are. For example, the mean (average) IQ score of our 100 students may be 118. However, not all students will have scored 118 on the IQ test. Rather, their scores will be spread out. Some IQ scores will be lower and others higher than the mean score of 118. Measures of spread help us to summarize how spread out these scores are. A number of statistics are available to describe this spread, including the *range*, *variance*, and *standard deviation*.

1.6.2 Inferential Statistics

Whereas descriptive statistics is the branch of statistics that involves organizing, displaying, and describing data, inferential statistics is the branch of statistics that involves drawing conclusions about a population based on information contained in a sample taken from that population. With inferential statistics, the researcher is trying to draw conclusions that extend beyond the immediate data. Thus, when conducting research, the researcher employs inferential statistics to serve two major purposes: (1) to infer from the sample data the attitudes/opinions of the population and (2) to aid the researcher in making a decision regarding whether an observed difference between groups (e.g., gender difference in problem-solving ability) is a meaningful/reliable difference or one that might have happened by chance alone. Thus, whereas descriptive statistics is used to simply describe what is going on in a data set, inferential statistics is used to infer from the data set to more general conclusions.