Mohammad Golshan, Ph.D.

Key Concepts in Research and Data Analysis

Key Concepts in Research and Data Analysis

Mohammad Golshan, Ph.D.

Key Concepts in

Research and Data Analysis

Mohammad Golshan, Ph.D.

(Assistant Professor of TEFL)

سرشناسه: گلشن، محمد ۱۳۵۵
Golshan, Mohammad پدید آور
عنوان و نام پدید آور: Key Concepts in Research and Data Analysis
مشخصات نشر: یزد، نخبگان فردا، ۱۴۰۲– ۲۰۲۴م.
مشخصات ظاهری: ۲۱ × ۱۴/۸ ۱۸۳ ص
شابک:978-600-8151-69-2
وضعیت فهرست نویسی: فیپا۱۵۰۰۰۰ تومان
یادداشت: فارسی انگلیسی
آوانویسی عنوان: کی
موضوع: روش تحقیق
موضوع: Dissertations, Academic -- Authorship
موضوع: تحقیق -- روش‌شناسی
موضوع: Research – Methodology
موضوع: Technical writing -- Science and technology
رده بندی کنگره: ۱۳۹۵ ٤پ۸گ/ LB2369
رده بندی دیویی: ۸۰۸/۰۲
شماره کتابشناسی ملی: 4497742

عنوان کتاب: Key Concepts in Research and Data Analysis
مؤلف: محمد گلشن
ناشر: نخبگان فردا
طراح جلد و ناظر فنی: محمد جعفر فلاح
نوبت و سال چاپ: اول۱۴۰۲
قطع و شمارگان: رقعی، ۱۰۰۰ نسخه
قیمت: ۲۵۰۰۰۰ تومان
شابک: 978-600-8151-69-2

انتشارات نخبگان فردا: ۰۹۱۳۳۵۲۹۱۴۹ www.nokhbeganfarda.com
مرکز پخش: تهران، خیابان انقلاب، خیابان فخر رازی، نبش خیابان لبافی نژاد، پلاک ۲۰۰، طبقه دوم، انتشارات الوند پویان
تلفن:۰۹۱۲۳۵٤۲۰۳٦

Dedicated to the memory of my mother:
the guiding light whose lessons of life bloom
forever in my heart.

Contents

Preface

The book *Key Concepts in Research and Data Analysis* is written to provide an accessible introduction to the fundamental concepts that underpin research and data analysis in various fields, particularly language teaching. If you are a university student, or simply someone interested in understanding the major concepts in research, this book can equip you with the essential knowledge needed to navigate the realm of research and data analysis confidently. Research and data analysis are powerful tools that can unlock new insights and drive meaningful changes in your approach. I would be grateful if you could write to me at the following address to share any ideas or comments that might help enhance the quality of future editions of the book.

Mohammad Golshan
Email: mohammadgolshann@gmail.com
Tel: +98 913 352 9149
Website: www.nokhbeganfarda.com

Acknowledgements

Special acknowledgements go to Dr. Mahdieh Sattar for checking and proofreading the manuscript before publication.

Unit 1

Science, Theory and Hypothesis

Why Research?

Research refers to the steps that are taken in order to collect and analyze information. It aims to increase our understanding of an issue. In simple terms, research is a systematic investigation to discover facts, revise and develop knowledge, and theories.

If you are a teacher or a person involved in education, you need to be informed about the changes and improvements in your field and therefore studying research enables you to be a good consumer of research. Studying research will give you a chance to judge the quality of the research conducted by others. Moreover, as a teacher you can examine what you are doing in your classes and find the answers to your questions alongside submitting your pedagogic practices to critical scrutiny. There are different ways through which you can achieve the above goals.

Library Research vs. Empirical Research

One way of finding the answers to your questions is by looking at what other people have said about a particular issue. This type of research is called **library research.** It is also known as **secondary or conceptual research.** It is valuable because we

should not ignore what other people have found out and repeat what they have done. There is an age-old proverb that says "Don't reinvent the wheel". This proverb implies that there is no need to waste time and effort creating something that already exists and works well. It's better to use existing solutions and build upon them instead of starting from scratch.

Another way of finding the answers to our questions is by carrying out an **empirical study** which requires collecting data and then drawing conclusions. It is based on our own data. This is what researchers mostly do when they test hypotheses to verify or reject theories or collect data to form new hypotheses and theories.

The word *empirical* in an *empirical study* has two key meanings: It can refer to a study that is based on observation and experiments. An empirical study is a study that relies on observable data collected through direct observation, experimentation, or measurement. This data is not based on opinions, beliefs, or theoretical models, but rather on concrete evidence gathered from the real world. For example, if you are researching the effectiveness of a teaching technique, you conduct a classroom study where you observe and measure the

effect of the technique on the learners' knowledge. Another meaning of *empirical* is related to verifiable evidence. An empirical study aims to generate findings that can be objectively verified and replicated. This means the study design, methods, and data analysis should be transparent and well-documented, allowing other researchers to repeat the study and obtain similar results. The main goal of an empirical study is to establish robust and generalizable conclusions that can be applied beyond the specific research context. This is the major way that moves science forward. Have you ever wondered what *science* means?

Definition of Science

Science refers to an approach to gathering knowledge. It has two goals. One goal is developing new theories and the other goal is testing the hypotheses that are deduced from theories. Therefore, the job of a scientist is to use existing theories for research, modify them or create new theories. Therefore, as was stated before, research can make an immeasurable contribution to science by testing already existing theories through formulating hypotheses and in some cases developing new theories. Here, in our definition of science, we frequently referred to the terms *theory* and *hypothesis*. Therefore, it is

time we focused on what these terms mean in the realm of research.

Definition of a Theory

A theory is an explanation that is based on evidence and has been tested over time. Theories are not mere guesses or opinions. They are based on evidence and careful thought, not just random ideas or beliefs. They can change and grow. As we learn more and gather more evidence, theories can be adjusted or even replaced by better ones. It is an ongoing process of discovery. They're powerful tools and can help us understand the world, make predictions, and solve problems. A theory describes the relationship among variables to explain and predict future occurrences. It establishes a cause-and-effect relationship between variables with the purpose of explaining and predicting phenomena. For example, the Germ Theory of Disease claims that germs can cause diseases. Evidence for this theory comes from the observation of microorganisms or germs, and the experiments that have demonstrated the role of germs in infection, and successful prevention of disease through hygiene and vaccination.

Definition of a Hypothesis

Hypotheses are used for development of a theory and for stating parts of an existing theory in a testable form. We can develop a hunch based on theory, past experience, observation and information gained from others. The hunch is changed into a hypothesis to be testable. Based on the findings of subsequent research, the hypothesis is verified or rejected and more hypothesis are formulated to continue building a cohesive theory. As to testing a theory, when we face a problem, we propose informal hypotheses that can be tested directly.

To clarify the meaning of a hypothesis, I cite an example from MacKey and Gass (2022). Suppose you are stuck in traffic. You naturally wonder why it's happening and come up with possible explanations like an accident or rush hour. Then, you try to confirm your guess by looking for clues, like seeing an accident, hearing a radio report, or checking traffic apps. If you find evidence, you can confirm your hypothesis. If not, you might conclude it's just regular traffic. We constantly ask questions, form hypotheses, and seek answers. Forming hypothesis is an essential part of our everyday lives. We do the same thing in research. However, most hypotheses in research are formed based on existing theories.

Unit 2

Research Types

Research Types

There are many different types of research, depending on the purpose, methodology, and design used. Research can be classified by goal: **fundamental research** and **applied research**. Fundamental research aims to expand knowledge and understanding in a specific area, without any immediate practical application in mind. It is often called **pure** or **basic** research. On the other hand, applied research focuses on finding solutions to real-world problems or developing practical applications of existing knowledge.

Research can also be classified by methodology. In this case we can have **quantitative research** that uses numerical data and statistical analysis to test hypotheses and draw conclusions. On the other hand, we have **qualitative research** that examines non-numerical data like interviews, observations, and texts to understand experiences, meanings, and perspective. There are some research types in which both quantitative and qualitative research are used together. We call these types of research **mixed methods research**.

Research can also be categorized by research design. In this classification, we have **experimental research** that tests a

specific hypothesis by manipulating variables and observing the results. **Observational research** collects data without manipulating variables, often to examine relationships between variables. **Descriptive research** describes or characterizes a particular population or phenomenon. There is a type of research called **correlational research** that investigates relationships between variables to see if they are related, but cannot establish cause-and-effect. We also have **survey research** in which we explore opinions, beliefs, attitudes, and perceptions of people.

With regard to time, a survey can be **cross-sectional** or **longitudinal**. In a **cross-sectional survey**, the researcher collects data from a sample at a specific point in time to understand the relationships or characteristics of a population while in a **longitudinal survey** study, the researcher studies a sample or population over an extended period. It gives the researchers the possibility of observing changes, trends, or patterns over time to understand the effects of variables across different stages.

We have other types of research such as **action research** that aims to solve specific problems in a particular setting, often

conducted by teachers in their language classes. **Case studies** deals with in-depth studies of individual cases or groups to gain insights into a particular phenomenon. **Ethnographic research** is a type of qualitative research that involves immersing the researcher in a specific cultural or social setting to observe and document the behaviors, interactions, and beliefs of individuals within that context. It often entails participant observation and in-depth interviews.

There is also one type of research called **meta-analysis**. Meta-analysis uses the statistical analysis of existing research studies to synthesize and summarize their findings. It aims to provide a comprehensive overview and draw conclusions by combining data from multiple studies. The researcher tries to decide which method or policy or technique is more effective based on a large number of studies which have been conducted.

You should keep in mind that these categories are not mutually exclusive. For example, a research project might be both qualitative and descriptive, or applied and experimental.

Unit 3

Variables

Definition of a Variable

A variable is a characteristic or quality that varies in people or objects. For example, age, gender, weight, etc. are variables. The opposite of a variable is a **constant**. A constant is a fixed value within a study. If all subjects in a study are eighth-graders, then grade level is a constant. In a study comparing the attitudes toward school of high school boys who plan professional careers with those high school boys who do not plan professional careers, high school boys is a constant.

Variable Types

Three important types of variables which are usually interesting in most of the studies include **independent variable** (also known as the treatment variable; experimental variable; manipulated variable), **dependent variable** (also known as outcome variable) and **moderator variable**. An **independent** variable is the variable that has a likely effect. It is the treatment that you do. It is usually used after the words such as *effect of,* and *impact of*. A **dependent** variable refers to learners' scores in post-tests and delayed posttests. We call it dependent because learners' scores in posttests and delayed posttests depend on the treatment. A **moderator** variable is the variable that moderates the effects of the independent variable

on the dependent variable. In the previous example of "the effects of two types of error correction methods on Iranian EFL learners' writing improvement", error correction is the independent variable that has two levels, i.e., direct and indirect and our dependent variable is learner's improvement which is measured through posttests or delayed posttests. If we assume that the effects of different types of error correction might be different for males and females, gender is a moderator variable which comes between the effects of the independent variable on the dependent variable. Gender as a moderator variable has two levels: male and female.

There are four other types of variables which are important in any type of research. They are important in the sense that if they are not taken into consideration by the researcher, the internal validity and external validity of the research will be affected. **Internal validity** means the extent to which the results of a study can be attributed to the treatment and **external validity** means the extent to which the findings can be generalized. These two important features of research are discussed in the Methodology. These important variables include **control variable**, **extraneous variable**, **confounding variable**, and **intervening variable**.

A **control variable** is the variable which the researcher holds constant or controls during an experiment, and is not part of an experiment. It is neither the independent nor the dependent variable, but it is important because it might have an effect on the results if it is not controlled. For example, if a researcher conducts an experiment only on boys and holds the gender of the participant constant, we say gender is a control variable because it has been controlled in the study.

Extraneous variables refer to any variable that you are not intentionally studying in your study but can threaten the internal validity of your results. They are variables that influence the outcome of an experiment, though you are not interested in their effects. For example, if a researcher who intends to investigate the effects of computers on language learning uses two classes, one of which full of boys and the other one full of girls, here the gender has not been controlled and since boys like computers more than girls, the results of the study might be due to the differences in the gender of the learners in each class rather than the use of computers. In such a situation, gender is an extraneous variable. As a researcher, if you do not control an extraneous variable, it becomes a **confounding** variable.

A confounding variable is a variable that obscures the effects of the independent variable on the dependent variable. If an experienced English teacher used a specific vocabulary teaching technique in his/her class and another instructor who had less experience did not use that technique, and students in the two classes were given achievement tests to see how many words they had learned, the independent variable which is the vocabulary teaching technique would be confounded. There is no way to determine if differences in vocabulary size between the two classes at the end of treatment were because of the experience of the teacher in teaching or the used vocabulary teaching technique.

An **intervening variable** which is also known as a **mediating variable** is a hypothetical variable that is used to explain causal links between other variables. Intervening variables cannot be observed in an experiment and because of this, they are called hypothetical. For example, there is a relationship between being rich and having a longer life span. But just because a person is rich doesn't mean that being rich will lead to a longer life. In fact, being rich results in more access to healthcare and better hospitals and these privileges result in a longer life. Therefore, other hypothetical variables are used to explain the

phenomenon. The intervening variables in the above example include more access to healthcare or good hospitals.

Variables can also be distinguished as **categorical** or **continuous**. When researchers classify subjects by sorting them into mutually exclusive groups, the attribute on which they base the classification is termed a categorical variable. Home language, county of residence, gender, and marital status are examples of categorical variables. The simplest type of categorical variable has only two mutually exclusive classes and is called a **dichotomous variable.** Male–female, and pass–fail are dichotomous variables.

When an attribute has an infinite number of values within a range, it is a **continuous variable.** As a child grows from 40 to 41 inches, he or she passes through an infinite number of heights. Height, weight, age, and achievement test scores are examples of continuous variables.

Now let's see what the meaning of a construct is . Constructs are abstractions that cannot be observed directly but are useful in interpreting empirical data and in theory building. For example, people can observe that individuals differ in what

they can learn and how quickly they can learn it. To account for this observation, scientists invented the construct called intelligence. They hypothesized that intelligence influences learning and that individuals differ in the extent to which they possess this trait. Other examples of constructs in educational research are motivation, reading readiness, anxiety, underachievement, creativity, and self-concept.

Unit 4

Research Questions and Hypotheses

Null vs. Alternative Hypothesis

Research questions are very important in research because after you make them, the reader finds the direction of your research and tries to see how you will answer them. After you pose your research questions, you formulate the hypotheses. Hypotheses are predictive statements and guesses about the outcome of a study. These predictions and guesses are temporary or tentative because you still wish to see if your guesses are correct or not and to find this out requires you to collect data and analyze it.

There are two main types of hypotheses: **Null hypothesis** and **alternative hypothesis**. The null hypothesis predicts that there will be no effect, no difference or no relationship between two variables. An alternative hypothesis predicts that there will be an effect, a difference or a relationship. It is called alternative because it is an alternative to the null hypothesis. To put it simply, if you state the null hypothesis in the affirmative form, it will be an alternative hypothesis. If we reject the null hypothesis, we can accept the alternative hypothesis.

There is also another type of hypothesis which is called **alternative directional hypothesis**. This type of hypothesis is posed when the researcher is a little confident about the likely

outcomes of the study and can make a specific prediction. Which hypothesis should you use? The answer is, if based on the literature review you can predict the difference or the relationship, you had better use an alternative directional hypothesis. If you cannot predict what will happen, use a null hypothesis.

a null hypothesis (H_0): There will be no difference between the experimental group and the traditional group.

an alternative hypothesis (H_1): There will be a difference between the experimental group and the traditional group.

an alternative directional hypothesis (H_{A-1}): The experimental group will outperform the traditional group.

Research questions without a moderator variable

To what extent do different language environments impact participants' writing performance?

Does a meaning-focused pre-writing task yield higher global quality scores on learners' writing than grammar-focused activities?

Research questions with a moderator variable

Does anxiety affect test performance and, if so, does it depend on test-taking experience?

Do Iranian EFL university students taught primarily by the CALL method perform better on tests of critical thinking than Iranian EFL university students taught primarily by the traditional method and, if so, does it vary with gender?

Experimental Study

H_0: There will be no statistically significant difference in the Japanese vocabulary test scores of beginning students taught with Total Physical Response and those taught with traditional methods.

H_1: There will be a statistically significant difference in the Japanese vocabulary test scores of beginning students taught with Total Physical Response and those taught with traditional methods.

H_{A-1}: There will be a statistically significant difference in the Japanese vocabulary test scores of beginning students taught with Total Physical Response and those taught with traditional

methods, with the students who had Total Physical Response instruction outscoring those who did not.

Correlational Study

H_0: There will be no statistically significant correlation between students' extroversion scores and their scores on a Spanish pronunciation test.

H_A: There will be a statistically significant correlation between students' extroversion scores and their scores on a Spanish pronunciation test.

H_{A-1}: There will be a statistically significant positive correlation between students' extroversion scores and their scores on a Spanish pronunciation test.

Unit 5

Literature Review

Introduction

One important reason why you should do a literature review for your study is to contextualize your research. Through reviewing the literature, you focus on the previous studies or a background theory which is relevant to your research project. When you review, describe, and synthesize the major studies which are related to the topic of your research, you will be in a position to show the relationship between your research and other studies that have been done in other parts of your country or the world. At the doctoral level, the literature review is very important because it should be as compressive as possible and at a state-of-the-art level. In summary, the literature review is usually done for the following purposes:

1) to search for a suitable problem area,
2) to learn about research which is already conducted into one or more aspects of the research problem,
3) to summarize the results of previous research to form a foundation on which to build one's own research,
4) to collect ideas on how to gather data,
5) to investigate methods of data analysis,
6) to study instrumentation which has been used,

7) to assess the success of the various research designs of the studies already undertaken.

Conceptual Literature and Research Literature

There are two types of literature which you should review as a researcher. The first type is **conceptual literature**. Conceptual literature is what the authorities and experts have written on the subject you are interested in, their opinions, their ideas, their theories and experiences. This type of literature is usually published in the form of books, articles and papers. The second one is **research literature** which provides you with the results of the research which has been conducted on the subject you are studying. This type of literature is often presented in the form of papers, articles and reports that are published in journals.

Alongside conceptual literature review which includes the key concepts, theories, experts' opinions and views, your literature review should report the major findings of the studies which were conducted in the past and were related to your topic, names of the researchers who conducted those studies, and the year when those studies were done or published.

You should go into details when you report and discuss the studies which are directly related to your own study. That is, you should give the readers information about the methodological approach that the researchers in those studies used, the way the data were gathered and also the method of data analysis which was employed. After mentioning all these details, you should provide critical comments on those studies.

Tell the reader which studies were the best studies; which studies had their own weaknesses, and also give readers reasons for your claims. Unfortunately, many students only present factual information about the studies and are afraid of making critical comments on them. In order to help you understand how to report in detail and how to provide critical comments, I give you an example from my own Ph.D. thesis. The part which is underlined is a critical comment on the report of this study.

Yang and Lyster (2010) undertook a quasi-experimental study to compare the effects of three different corrective feedback treatments on 72 Chinese learners' use of regular and irregular past tense. They assigned three intact classes to a prompt group, a recast group and a control group. The

students were engaged in form-focused activities that required the use of the target forms. The results of the pre-test, immediate post-tests and delayed post-tests that measured the accurate use of the target forms in oral and written production showed significant gains by the prompt group on eight measures, the recast group on four and the control group on three. The findings revealed that prompts were more effective than recasts in the correct use of regular past tense forms, whereas prompts and recasts were equally effective in improving accuracy in the use of irregular past tense forms.

Although this study showed that prompts are more effective than recasts, the findings should be interpreted with caution. First of all, the recasts in this study were not distinguished in terms of being explicit or implicit. Second, the operationalization of prompts included metalinguistic clue, repetition, clarification requests, and elicitations. Therefore, generalizing the findings to all types of prompts and recasts can be problematic.

As you can see in the above paragraph, the names of the researchers, the time the results of their study were published, their methodological approach and also the thesis writer's

critical comments were provided. In summary, when you report a study, try to answer the following questions:

- Who conducted the study?
- What was the purpose of the study?
- Where was the study done?
- Who were the participants in the study?
- How was the study conducted?
- What was the procedure of data collection and analysis?
- When was the study published?
- What were the findings?

When you want to have critical comments on the studies, ask the following questions:

- Did the researcher or researchers state the research problem clearly?
- Did they define the variables clearly?
- Did they use a sound design to answer the research question?
- Were appropriate research instruments used in the study?
- Was the statistical technique the best one?
- Are the researchers' conclusions and implications logical, based on the results of the study?

Position Towards Previous Research

When you review and critique previous research, you should also show your position, or stance, in relation to those studies. You should show to what extent you believe the findings are true. In everyday conversation, when you say *perhaps*, *I guess*, *maybe*, you are trying to reduce the strength of your utterance. The following situations are the cases when you show your position:

1) You wish to show that you do not have full commitment to a proposition.
The findings might/perhaps_____ It is possible that the researchers' findings _____.

2) You wish to show your certainty in something.
In fact, _____. It is clear that_____.

3) You wish to show your attitude.
Surprisingly _____, Unfortunately, _____.

4) You wish to involve the reader in your thinking process.
Consider_____, Note that_____, You can see that _____.

A question that you might ask is “From where should I start?” The first thing that you need to do is to identify the key authors and journals related to your field. Then start reading the latest articles on your topic. Use *Google Scholar*, *Tables of Contents from key journals*, and *reference lists of the articles*, *books and chapters*.

After you find the related articles and books about your topic, critically read the literature. Focus on strengths and weaknesses of previous studies, the methodology that has been used, the measurement instruments, data analysis techniques, and their results.

Make a preliminary outline for the Literature Review chapter of your thesis by deciding about the sections and subsections of the literature review. Do not worry because you can change this outline later. Limit the scope of the review to your own topic. Start writing the review. Avoid plagiarizing. Use the review to lead to the gap in the literature, your study and research questions.

There are several ways you can report on previous studies. One way is that you place the author in subject position in the sentence. This method of reporting is called *central reporting*.

E.g., Fanselow (1977) studied the error correction patterns of 11 experienced teachers. He found that teachers' CF did not go beyond showing the students whether their answers were correct or wrong.

Another way is to give less focus to the author and put his/her name in brackets at the end of the relevant statement. This method is called *non-central reporting*.
E.g., It has been shown that teachers' CF did not go beyond showing the students whether their answers were correct or wrong (Fanselow, 1977).

The third way is the results of a study are presented with less focus on the author or the actual study and no 'reporting verbs' such as 'claim' or 'show' are used. This is called *non-reporting*.

E.g., Teachers hold beliefs about teaching and construct their own personal theories of teaching (Woods, 1996).

Vocabulary focus in literature review

The common verbs for reporting on literature review include:

point out	dismiss	propose
add	dispute	question
affirm	doubt	recommend
argue	explain	report
assert	identify	state
believe	indicate	suggest
challenge	maintain	support
claim	observe	think
describe	present	urge

Unit 6

Research Methodology

Introduction

The Methodology is sometimes labeled *Method* is one of the most important aspects of research and some researchers refer to it as the skeleton of the study. Methodology deals with the following questions:

1) Who were the participants you selected for your study? (characteristics of the participants)
2) How were the participants selected? (type of sampling)
3) Where did the study take place? (setting)
4) What variables are involved in the study and how will they be manipulated? (design)
5) What did you use for treatment and data collection? (materials and instruments)
6) What did you do? (procedure)
7) How were the data collected and analyzed? (data analysis)

The methodology part of a study is important because it enables the readers to understand what you did, allows them to evaluate the appropriateness of your methodology, and enables them to replicate your study in the future if they wish to do so. The ability to replicate a study is an important criterion which

is used to judge the quality of the methodology section of an article. Therefore, in the Methodology section of a research paper or article you have the following sections.

1) Participants
2) Sampling
3) Design
4) Materials and Instruments
5) Procedure
6) Data Analyses

These sections are not fixed in order and there is variation in the order of these sections in different theses. Some researchers prefer to start their methodology with Design while some prefer to begin with Participants. Whatever order you follow, keep in mind that **Participants**, **Procedure** and **Data Analyses** are the essential sections of a Methodology chapter.

In some theses, you might come across a section which is labeled **Setting**. Since I did not come across this section in many articles and theses, I will not explain it as a separate section. and just elaborate on the function of this section.

Consult your supervisor about whether you include this section before the Participant section or not.

Participants

In this section of the Methodology, you should describe the participants/subjects or the objects of the study from which you have collected the data. Try to provide as many details as needed about the participants. You should also discuss the rationale used for selecting the participants so that the readers can judge whether the data that you have gathered from these participants are valid for the purpose of the study or not. The essential pieces of information that you need to provide concerning the participants of your study include:

1) **the number of participants**

Twenty-two participants

88 Persian-speaking fourth-graders

The sample consisted of a total of **6** intact classes.

2) **their ages**

Twenty-two advanced EFL participants **aged 12–25**

88 Persian-speaking fourth-graders, **aged between 9 and 11.6** years (M = 9.2 years, SD = 0.52),

Sixth-grade students between **10 and 12 years of age** (n = 29; male = 12; female = 17)

The participants were all adults and **ranged in age from 20 to 50**.

3) **their nationalities**

One hundred and eight 19-year-old **Iranian** EFL learners were recruited for this study.

The sample included 60 male **Iranian** students between the ages of 20 and 30 years (*M* = 24.2, *SD* = 2.1).

4) **their gender**

Twenty-two participants (**6 males, 16 females**)

Sixth-grade students between 10 and 12 years of age (n = 29; **male = 12; female = 17**)

5) **their educational level**

400 English teachers **holding BA and MA degrees in English** literature participated in the study.

The students were all female and varied greatly in terms of educational background, ranging from **freshmen to seniors**.

The participants were 60 elementary learners who were randomly selected from among 100 learners, who were **placed**

at intermediate level by an English language institute in Iran based on placement tests that they had taken.

6) **how they were assigned into groups**

Four fourth-grade, 4 fifth-grade, and 4 sixth-grade intact classes (n = 318 students) were **randomly assigned to experimental** and control conditions

7) **how much class time the participants had completed before the study started**

All the learners had learned English in an instructed setting **for about one year**.

8) **their language background**

They had received between **6 months of English instruction** either at the same language institute or in high school.

The participants had **never visited an English-speaking country**.

9) **their first language**

Participants were **native speakers of Persian** in four intact fourth-semester language classes.

The sample included 120 male Iranian students between the ages of 20 and 40 years, all of whom were **native speakers of Persian** and had at least 12 years of education.

As far as the objects are concerned, for example books, you should mention why those books were selected and all the things related to them. One of the questions that graduate students frequently ask is how many participants they should select for their study? What is the best sample size? The important point is that there is not an exact rule based on which we can set the optimal sample size. I try to offer some guidelines here based on my review of the literature.

One way to get to an optimal sample size is through a rough figure or method of calculation, based on practical experience. As far as a survey is concerned, between one percent and ten percent of the population is an appropriate size with a minimum of about 100 participants, but the size of the sample can be smaller if scientific sampling rules are used in sampling. Because in opinion polls researchers use scientific rules of sampling, a very small sample such as 0.1 percent of the population is used as a sample. Researchers believe that at least 30 participants are required for correlational studies, at least 15

participants are required for each group in experimental groups and at least 100 participants for factor analysis and multivariate procedures (Mackey & Gass, 2005).

There are further rules about sample size. For example, the larger the sample a researcher selects, the smaller the magnitude of the sampling error will be and the sample is more likely to be the representative of the population. For survey studies we should have larger samples than those of the experimental studies because those who participate in the survey are volunteers. Besides, if survey questionnaires are mailed to the participants, the percentage of responses may be as low as 20 percent or 30 percent. Therefore, at the beginning, a large sample should be selected. When the researcher wishes to subdivide the samples into smaller groups, a large enough sample should be selected by the researcher so that subgroups are of adequate size later.

Unit 7

Different Types of Sampling

Sampling

In research, we usually deal with two terms: **population** and **sample**. Population refers to any group of individuals who are similar in some respects. Therefore, a group will be population when individuals have at least one common characteristic (e.g., the population of high school students because they all study at high school). Since most of the time the population size is big and we cannot investigate the whole population, we get a **sample** from the population and to solve the problem of diversity, we limit the population to a specific group to which we wish to generalize our findings.

Different Types of Sampling

Sampling can be divided into two groups in general: **probability** and **non-probability sampling**. You had better use probability sampling as much as possible. Although probability sampling is usually considered to be a scientifically sound sampling, it is expensive and applied linguists can seldom use it. Researchers in applied linguistics usually use non probability sampling.

Probability Sampling

Random Sampling

As the name suggests, a sample is selected randomly from the population through a random numbers table or computer-generated list. The point is that random samples are usually more representative than non-random samples.

Stratified Random Sampling

In this type of sampling, which is a combination of sampling and categorization, the population is divided into strata (i.e., groups) and then a random sample of proportion is selected. The first thing that the researcher needs to do is to identify some of the parameters of the population which are important in his/her research, such as gender (i.e., males and females) and then he/she should think of selecting the participants randomly from those parameters.

Systematic Sampling

In this type of sampling we select every n^{th} number of the target group. For example, if you wish to select 100 names in a telephone directory for a survey, and there are 20000 listings, you choose the first name randomly from one of the pages and

then go on to select every 200th name until a sample of 100 names was selected.

Cluster Sampling

This type of sampling is also called *area sampling* and when the target population is widely dispersed and we do not have a list of the members of the population, we randomly select units of population and then examine people from those units. For example, a researcher wishes to survey high school students. From all the high school students in a city, he chooses some high schools randomly and then studies them.

Stage Sampling

This type of sampling is an extension of cluster sampling. The researcher selects the sample in stages - taking samples from samples. For example, the researcher who has limited the choice to schools from a particular area may then randomly choose only a limited number of schools from those available - and then only a limited number of students are chosen at random from within the selected schools.

Non-Probability Sampling

Quota Sampling

Quota sampling is similar to stratified random sampling but the participants are not selected randomly. Similar to stratified sampling, the population is divided up into groups with similar characteristics (for example, males and females), and then members are selected randomly from within these groups, but the numbers selected are in proportion to their occurrence within the whole population.

Dimensional Sampling

Dimensional sampling is an extension to quota sampling. In this type of sampling the researcher takes into consideration several characteristics, such as gender, age, income, residence and education. The researcher tries to ensure that there is at least one person in the study representing each of the chosen characteristics. For example, out of 10 people, the researcher makes sure he/she has interviewed 2 people that are a certain gender, 2 people that are from a certain age group and 2 who have an income between a range. To put it simply, this is a simplification of quota sampling employed to reduce sample size. The researcher identifies various factors of interest within the population (for example, it may be important to include the

responses of people of different ages) and then ensures that the sample includes respondents from each of the groups thus identified.

Snowball Sampling

This type of sampling is used when access to population members is difficult or they are hard to identify. The researcher uses a few members of the population and through those members new members are identified and again from the new members, he/she tries to identify and find new members until a sufficient sample size is reached. For example, you would like to conduct a study on drug addicts. It is very difficult to identify these people. So you ask a few of the drug addicts you know to introduce those addicts whom they know. Then you refer to the newly identified drug addicts and ask about more addicts that they know and this process goes on. This type of sample involves chain reaction.

Convenience Sampling

Convenience sampling, which is sometimes called *opportunity sampling*, or *availability sampling* is the most common type of sampling in L2 research and is usually used when the participants possess certain key characteristics that are related

to the purpose of the investigation. It consists of those persons available for the study. For example, a teacher working in a language school has access to the students in that school; therefore, he/she chooses some classes in that school and conducts an experiment in those classes.

Purposeful Sampling

Purposeful sampling which is also called *purposive sampling* is mostly used in qualitative studies. For example, the researcher focuses on a small number of people who have the characteristics that the researcher is interested in. Suppose that a researcher wishes to study the attitudes of the managers who have been dismissed from their jobs. He selects four managers who have been recently dismissed from their jobs because those managers serve the purpose of the researcher.

Unit 8

Research Design

Design

There are various types of designs and as a researcher you should select the best type for your study. The major classes of quantitative research designs are **pre-experimental**, **quasi-experimental**, **ex post facto**, and **true experimenta**l. These classes of designs have more specific designs within them. Since pre-experimental design is seldom used nowadays, our focus will be on the other types of designs. Quasi-experimental design can be divided into three specific designs, which include **comparison groups designs**, **time series designs** and **equivalent time samples designs**. Ex post fact design can be divided into **criterion group designs** and **correlation designs**. True experimental design can be divided into **post-test only control group design** and **pre-test post-test control group design**.

In another classification, designs may also be categorized as **correlational designs**, **experimental designs and quasi experimental designs**, **repeated measures design**, **factorial designs**, **time series designs**, and **one-shot case study** (see Mackey & Gass, 2005 for a review). I will elaborate on the most common types of designs which are used in quantitative studies.

Survey Design

A survey study provides a quantitative description of beliefs, attitudes, or opinions of a population by studying a sample that is selected out of that population. The main goal of a survey is generalizing from a sample to a population. It can be either **cross-sectional** or **longitudinal**. In a cross-sectional survey design, the data are collected at one point in time. For example, when you administer a questionnaire to university students in the middle of a course to see how satisfied they are with the course. In a longitudinal survey design, data are collected over time. For example, when you examine the anxiety trend of the students from the beginning to the end of the term. You administer the questionnaire every two months and then examine the trend over time to see whether anxiety is decreasing or increasing. A survey study employs questionnaires or interviews for data collection.

True Experimental Design

We compare groups based on the effect of an intervention and the results of the tests. In this design, the researcher wishes to see if a specific treatment influences an outcome. The researcher tests this by providing a specific treatment to one group while another group which is usually called a control

group does not receive treatment. Then the researcher determines how both groups scored on an outcome.

Quasi-Experimental Design

It has all the characteristics of a true experimental design but does not include random assignment of the participants to the groups. In simple terms, if the sample is selected based on convenience sampling and intact classes are used, the study will be quasi-experimental.

Repeated Measures Design

Repeated measures design is also known as **within-group** design. In a within-group design, the behavior of a single individual or a group over time is studied. Here the experimenter provides and stops a treatment at different times in the experiment to determine its impact. If you measure a single group's level of motivation before a course starts and then administer the same questionnaire at the end of the course, you are using a repeated measures design because the pretest and posttest of a single group are being compared here.

Time Series Design

Time series design is a design in which repeated observations over a set period of time take place and the participants serve as their own control. For example, you want to find out whether introducing a midterm to the course can affect the students' attention to the details in the course materials. At the end of every week before the midterm, you give the students a quiz and you collect data on their attention to various details of the course materials. After that, you introduce the midterm and then again give them quizzes to collect data on whether they paid more attention to the details after the midterm exam. If you compare the students' average scores on quizzes after the midterm with those of before the midterm, you are using a time series design. This design is frequently used with small groups of learners.

Ex Post Facto Design

In ex post facto design, the researcher investigates the independent variable which has already taken place. *Ex* means "experiment"; *post* means "after"; and *facto* means "fact". Therefore, ex post facto means "after the fact". In other words, two or more groups are compared although no experiment is

going on. Ex post facto design can be classified as **criterion groups** design and **correlational design**.

Criterion Groups Design

Suppose that you wish to see whether female students are better at listening comprehension than male students. Here the variable of gender as an independent variable is not something that you can manipulate. People are already born male or female. This type of design is called criterion groups design. Investigating left-handed and right-handed people, or comparing Iranian and English teachers are examples of an ex post facto design in general and criterion groups design in particular.

Correlational Design

Correlational design is also known as **associational design** in the literature. This design is used to test a relationship between or among variables or to make predictions. If there is a relationship between two variables, we can often predict the likelihood of the presence of one from the presence of the other(s). If there is a negative correlation between age and memory, it suggests that as people get older, their memory power decreases.

Mixed Methods Design

In mixed methods design you combine both quantitative and qualitative data. In this case, just mentioning that your design is mixed methods does not suffice, and you need to elaborate on the specific mixed methods design that you used. Specific mixed methods designs include 1) **convergent parallel mixed methods design**, 2) **explanatory sequential mixed methods design**, and 3) **exploratory sequential mixed methods design**.

In convergent parallel mixed methods design, you collect qualitative and quantitative data and analyze them separately and then compare them to see whether the findings of each confirm or disconfirm the other one. In explanatory sequential mixed methods design, you collect quantitative data in the first phase and in the second phase use the results of the quantitative data to build the second phase which is the qualitative phase. For example, based on the results of a questionnaire data, certain participants are selected for an interview. Exploratory sequential mixed methods design is exactly the opposite of explanatory sequential mixed methods design. In this type of design the researcher first collects qualitative data and then based on the qualitative data, he/she moves to collect

quantitative data. For example, he/she interviews people and then designs a questionnaire based on the interview.

Two important points you need to consider

Designs vary based on whether they have a control group or not. Whenever you wish to find whether treatment has any effect on the dependent variable, there is a need for a control group. The control group takes the same pretest and posttest as does the experimental group, but does not have the same treatment between pretests and posttests.

If you want to find out whether music has any effect on learners who listen to music in class, you need to have a control group that does not receive music, but is matched to the other group in all other respects. In this case the design is a **control groups design**. However if you want to see whether 2 hours of listening to music is better than 4 hours of listening, there is no need for a control group. You just compare two groups who are exposed to two different amounts of music to see which amount is more beneficial. In this case the design is a **comparison groups design**. Since true experimental or quasi-experimental designs use either control group design or comparison group design, stating whether you had a control

group or a comparison group is an important part of the design section.

The second point is in experimental studies, the researchers might use a pretest, but they certainly use an **immediate posttest** or they might use a **delayed posttest**. When the sample is selected on a completely random basis, the researchers may avoid giving a pretest because a pretest might alert the participants to what the treatment is about. However, the main problem is that a researcher cannot be sure if there is initial comparability of groups. To solve this problem, researchers give a pretest. With regard to posttest, some researchers administer only the immediate posttest while other who are interested to know whether the treatment is long-lasting or not give a delayed posttest one, two or three months later. Therefore, stating whether a pretest, posttest, delayed posttest design was used is another important aspect of describing the research design.

E.g., A pre-test-post-test control group design with 3 Iranian Advanced EFL classes which were randomly assigned to one of the 2 experimental groups and one control.

In sum, a consistent classification of designs has not been made in the literature and different books provide different classifications. However, one thing is clear. You should first know whether your study is experimental, quasi-experimental, associational, survey, descriptive, etc., and in the second step, you should report in detail whether a control group was included or not, whether a pretest was administered, whether a delayed posttest was given and also enough information about your independent and dependent variables.

A checklist provided by Mackey and Gass (2005, p. 158-159) helps you to have a better grasp of the questions which should be answered while designing a study.

- Are your groups matched for proficiency?
- If you are using a particular type of task (e.g., listening), are your groups matched for (listening) abilities?
- Are your participants randomized?
- If intact classes are used, are their treatments randomly assigned?
- Are your variables clear and well described?
- Do you have a control group?

- Are control groups and experimental groups matched for everything but the specific treatment (including the time spent on the control and experimental tasks)?
- Have you described your control and experimental groups?
- Do you have a pretest?
- If you are testing development, do you have a posttest or even multiple posttests?
- If using a repeated-measures design, are the treatments counterbalanced?

Unit 9

Materials and Instruments in Research

Materials and Instruments

Materials

In some research designs, the researcher uses one treatment or several treatments. He/she manipulates an independent variable to see its effect on a dependent variable. The treatment may involve some techniques. For example, you want to see if using gestures can have any effect on language learning. The treatment can also consist of some types of materials. The materials may be presented to the participants in the form of written, audio, or visual information. For example, when you want to see the effect of reading simplified stories vs. unsimplified stories on language learning, you use treatment materials. You should be careful not to confuse treatment materials with data collection instruments. In the following section the instruments will be explained.

Instruments (Instrumentation)

This section is sometimes labeled **Instrumentation**. Let's see what the difference between these two terms is. Instruments relate to the devices that are used to collect the data. Instrumentation is related to a) the tools or devices by which researchers try to measure variables in a study, 2) instrument design, selection, construction, and assessment, and 3) the

conditions under which the designated instruments are administered. So the name you select as the heading depends on the purpose and nature of this section.

Instruments are usually in the form of questionnaires or tests, and responses can be gathered via paper-and-pencil tests, computer-administered tests, video cameras, or audiotape recorders. In the Instruments section, you should describe your tests and questionnaires in detail. You must provide the following information in the Instruments section

1) **the construct that the instrument measures**
   E.g. Two questionnaires containing Likert-scale and open-ended response items were created by the researchers to **elicit the learners' perceptions about computer-assisted learning**.

2) **whether the instrument was timed or untimed**
   E.g., A **timed** grammaticality judgment consisting of 24 multiple-choice tests was used for the purpose of the study.

3) **format of the instrument and items in the instrument**

E.g. The questionnaire contained a total of 20 items using a **5-point Likert scale** ranging from *strongly disagree* to *strongly agree*.

4) **what the participants were expected to do**

E.g., Respondents were required to **mark their responses** on a five-point Likert scale ranging from *strongly agree* (1) to *strongly disagree* (5).

5) **number of the items in instrument**

E.g., The questionnaire comprised **50** statements.

6) **developer or origin of the instrument**

E.g., Each teacher completed a 20-item survey Questionnaire **developed by** _____to examine her or his opinions about_____.

7) **whether the instrument was adopted or adapted**

E.g., A 36-item survey **adapted** from_____Scale developed by _____was administered to the students to measure their_____.

8) **whether the instrument has different sections or subscales**

   E.g., The second questionnaire included two sections: (a) **a scale assessing** _____ (adapted from _____); and (b) **a scale assessing the participants'** _____.

9) **whether the instrument was piloted before use**

   E.g., The questionnaire (see Appendix A) was **pilot-tested** with Iranian EFL teachers and then revised

10) **the reason you selected one test among the other existing instruments**

    E.g., For the purpose of this study, recognition tests were not appropriate **as participants were adults with diverse backgrounds that existing tests, originally devised for younger participants may not have been suitable.**

11) **whether the instrument that was translated was also back translated**

    E.g., All the items in the student's questionnaire were initially written in English and translated into Persian. **Back-translation** was performed to check the accuracy

of the Persian version of the questionnaire, which was then altered as required.

A **back-translation** is a procedure through which you take a questionnaire that has been translated into another language and translate it back into the original language and then compare the two. Back-translation is a useful tool because of sensitive translation problems across cultures and languages. The questionnaires need to be translated into the native language of the learners if the learners are not proficient enough to understand English. The translations should then be back translated and be approved by experts.

12) **whether necessary steps to ensure maximum reliability and validity are taken**

E.g., The reliability of the test was estimated using **Cronbach's alpha**. The reliability coefficient for 20 items in pretest produced an alpha of .88 and alphas of 0.86 and 0.89 for the posttest 1 and posttest 2 respectively.

What is very important in this section is that you assure the readers that your instruments had an acceptable level of **reliability** and **validity.** These terms will be discussed in the

next units. You can open a separate subsection under the Instruments section and explain the reliability and validity of the Instruments or you can incorporate the information into the Instrument section without having new headings.

Procedure

In the Procedure section, you should provide a detailed explanation of how the complete study was done. In some studies, the data-collection subsection and this section are the same. You should provide enough information for other researchers so that they can replicate the study if they wish to do so in the future. Suppose that you wish to examine the effects of two types of teaching techniques on Iranian university students' knowledge of grammar. Your independent variable is the teaching method which has two levels in your study (deductive and inductive) and your dependent variable is the scores on a test of grammar given as the pretest and the posttest. So, in the Procedure section you must specify:

1) How long did the treatment last? (how many sessions)
2) How long was each session?
3) When was the pretest given?

4) How much time was given to the learners to answer the test?
5) In what format was the test?
6) When was the posttest given?
7) When was the delayed posttest (if any) given?

Before Moving to the next section, I wish to highlight the role that **sampling**, **design**, procedure and **instruments** play in enhancing the quality of your study. In Chapter one, I mentioned that you should attempt to control the variables that threaten external and internal validity of your study. Here, I go into more details and explain the factors

Unit 10

Reliability

Reliability

Sometimes when you look at a research article, you see the phrases such as **reliability coefficient**, **Cronbach's alpha**, **KR20 reliability coefficient**, etc. These terms might seem confusing, but they are simpler than they might seem. Reliability means consistency. Suppose that you call a friend a reliable friend. It means you can rely on him now, tomorrow or next year. In fact your friend is consistent in helping you across different times and you can depend on him/her at all times. Like a friend, a test must also be consistent in the results that it produces from Time 1 to Time 2, from rater 1 to rater 2 or from item 1 to item 2.

In journal articles reliability takes three basic forms: 1) the extent to which a test is consistent across repeated testings, 2) the extent to which the individual items go together to make up a test that measures the same underlying characteristic, and 3) the extent to which raters who have rated the test have been consistent. These three basic forms are called: **test-retest reliability**, **internal consistency,** and **inter-rater reliability**. The numerical index or the number that shows the amount of reliability is called **reliability coefficient.** Each of these will be explained below.

Test-Retest Reliability

Sometimes a researcher measures a group of people twice by using the same measurement instrument or test and then correlates the scores of the test at Time 1 to the results of the same test at Time 2. He does so because he or she wants to see if the results of the two testing times are correlated or not. In other words, he wants to show how consistent or stable the results of the test are over time. If a person gets a high score at Time 1 on anxiety, he is expected to get a high score at Time 2, too. In fact, if you are a very anxious person (high score), you normally have to be a very anxious type of person tomorrow, next week, next month, or any other time too. If a test or a questionnaire shows you as a very anxious person at Time 1 and a relaxed person at Time 2, that test or questionnaire does not have **test-retest reliability**.

As I told you above, the numerical index or the number that shows the amount of reliability is called **reliability coefficient**. Like other forms of reliability, test-retest reliability coefficient ranges from 0 to 1. It will not go below 0 or over 1. The closer the coefficient is to 1, for example .90, the higher the reliability is. Be careful that when you report test-retest reliability, you must mention the length of time between the testing times.

Besides, the time interval should be neither too short nor too long. If you get a high correlation coefficient or high reliability coefficient when the time interval is long, you should be happy that your test is a good one with regard to test-retest reliability.

Internal Consistency

A test or an instrument consists of several items or individual questions. For example, in a test of motivation several questions measure a person's level of motivation. All the questions and items should measure motivation and hang together. In other words, if we have 10 items in a questionnaire and 9 of them are related to motivation and one item is not, that one item should be removed from the test to increase the internal consistency of a test. If a person who is highly motivated, gets a high score on Questions 1, 2, 3, 4,5,6, 8, 9, 10 in motivation questionnaire, but gets a low score on Question 7 while he was supposed to get a high score on this question (I emphasize "He was expected to get a high score on this item too but he didn't get"), you can naturally conclude that this item is not related to other items in a questionnaire and must, therefore, be deleted.

To calculate internal consistency, a test or questionnaire is given to a single group of individuals at a single time. After all responses have been scored, and entered into SPSS, internal consistency can be calculated easily. Like other forms of reliability, internal consistency coefficient also ranges from 0 to 1. The closer the coefficient is to 1, for example .90, .91, etc, the higher the reliability is. Internal consistency can mainly be calculated through the following statistical techniques: 1) **split-half reliability**, 2) **Kuder-Richardson 20/ Kuder-Richardson 21,** and 3) **Cronbach's alpha**. Each will be explained briefly below.

Split-Half Reliability

One type of internal reliability is **split-half reliability**. In this type of reliability, each examinee's performance is split into two halves. First the total score of odd-numbered questions is calculated and then the total score of the even numbered items is obtained, and finally each person's total score on each half is correlated through a formula which is called **spearman-brown formula**.

Kuder-Richardson 20/ Kuder-Richardson 21

The second way to calculate internal consistency is Kuder-Richardson 20. In this approach the test is given to the

individuals once. The results are better than split-half. Kuder-Richardson 21 is also used to calculate internal consistency. It is easier to calculate and before computers were available for calculating, KR21 was more popular than KR 20. However, nowadays KR20 produces better results and is reported more frequently than KR21 in journal articles.

Cronbach's Alpha

The third method to calculate reliability is **Coefficient alpha or Cronbach's alpha**. Cronbach's alpha has more uses than KR20. KR20 requires the items which are scored either 0 or 1, but Cronbach's alpha can be used with any range of scoring. For example, for a test in which you have used a wide range of values for scoring such as 1, 2, 3, 4, you can use **Cronbach's alpha**. In fact, since questionnaires usually use 5 or 7 point Likert scaling, Cronbach's alpha should be used as a measure of internal consistency.

Inter-Rater Reliability

If a test is scored by a rater, you expect the second rater to give the same or almost the same score to the test. If the score is consistent from rater 1 to rater 2, and both give the test an almost similar score, the test is said to have inter-rater reliability. There are five popular procedures which researchers

usually use to calculate inter-rater reliability. They include 1) percentage-agreement measure, 2) Pearson's correlation, 3) Kendall's coefficient of concordance, 4) Cohen's Kappa and the 5) intra class correlation. Each will be discussed briefly below.

Percentage-Agreement Measure

Percentage-agreement measure refers to percentage of the occasions in which raters agree in their ratings of a test. To calculate this type of reliability, the number of the occasions in which the raters agree in the ratings are divided by the number of agreements plus disagreements and the result is multiplied by 100. For example, if an essay is scored by rater 1 and he finds 10 verbs as errors and the second rater only considers 8 of those verbs as errors, we have 8 cases of agreement and 2 cases of disagreement. Therefore, we divide 8 by 8+2 which become 8/10 and then multiply it by 100 .The result will be an inter rater reliability of 80 percent.

Pearson's Product–Moment Correlation

In percentage-agreement measure, we mostly deal with categorical data. That is to say, we say whether something is correct or incorrect. In **Pearson's product–moment correlation**, we deal with only raw scores. For example, we

have 20 papers or essays. Two teachers or researchers rate each paper or essay independently and assign a score, for example on scale of 1 to 10, to the 20 papers. Then the scores of each teacher are correlated to scores of other teacher to see if there is a positive correlation between them. After calculating the correlation, inter-rater reliability coefficient is obtained. The closer the correlation is to 1, the higher the inter-rater reliability is.

Kendall's Coefficient of Concordance

If the raters rank the data, then **Kendall's coefficient of concordance** is the best option for inter-rater reliability. Suppose that there are five people and each rater ranks each essay with regard to quality from 1 to 5. In order to find whether the raters have ranked the five essays in the same way, this procedure is used. The closer Kendall's coefficient to 1, the higher the amount of agreement between the rankings of the raters.

Cohen's Kappa

Like Kendall's coefficient of concordance, **Cohen's Kappa** is also used to calculate inter-rater reliability. However, Cohen's Kappa is used when the data are categorical. For example, suppose that two or more raters have to decide whether some

errors are grammatical errors or lexical errors. If all the raters agree that a certain error is grammatical, and there is total agreement between them for all the errors, inter-rater reliability or Kappa will be +1.

Intraclass Correlation Coefficient

Intraclass correlation is used to evaluate the level of agreement between raters in measurements. This method is better than ordinary correlation as more than 2 raters can be included. Intraclass correlation coefficient shows concordance and a coefficient of 1 is perfect agreement and 0 is no agreement at all.

Standard Error of Measurement

Sometimes researchers refer to **standard error of measurement** (SEM) as evidence for reliability. For example, suppose that a test is given to a group of students, and one of the students obtains 140. If standard error of measurement related to scores is equal to 5, we can think of an interval for that student. 140-5 to 140+5. In other words, if a test is given to the same group again, we assume that student is likely to get a score between135 to 145. This confidence ban is related to reliability in a reverse manner. The higher the reliability, the smaller the interval and the lower the reliability, the longer the

interval. Another difference between standard error of measurement and reliability is that reliability ranges from 0 to 1, but SEM depends on the data, scores and measurement units.

Important Points About Reliability

There are some points that need to be considered. First of all, if there is a high coefficient of stability, it does not mean that the internal consistency is high too. Moreover, a high value for split-half reliability does not imply that K-R 20 is also high. Therefore do not generalize from one method of reliability to all methods. You had better use different methods of reliability in one study.

The second point is that reliability is the property of the data, and not the measuring instrument. It might vary across groups that are different in gender, age, level and you had better not cite reliability coefficients obtained by previous researchers, but collect reliability evidence for your own investigation. It fact, reliability should be reestablished for any study.

The third point is that reliability coefficient is just an estimate of consistency. Therefore, use the word estimated with the word reliability.

The last point is that since the element of time at the time of administering a test increases or decreases the estimates of reliability, you must mention under which conditions (i.e., timed test vs untimed test) the data were collected. In other words, different estimates of internal consistency will be high when a test is given under great time pressure.

Unit 11

Validity

Validity

As was mentioned in the previous chapter, reliability refers to consistency and stability of the scores. In this chapter, another important concept which is called **validity** is introduced. Validity refers to the degree to which a test or questionnaire measures what it is supposed to measure. In fact, validity refers to the *appropriateness, correctness, meaningfulness,* and *usefulness* of the specific *inferences* researchers make based on the data they collect from a test or a questionnaire.

A test can be reliable but not valid. Imagine that you have an English grammar test which is reliable with regard to test-rest reliability, and internal consistency. That is to say it produces consistent results. What happens if you use that test to measure students' ability to do critical thinking? Is a reliable English grammar test a valid indicator of critical thinking ability? Definitely not. Therefore, we can conclude that reliability is necessary but not sufficient. But if a test score is valid, it is very likely that it is reliable. If a test is a valid measure of mathematics knowledge, a person who has a good knowledge of math is likely to get almost consistent high score on the test. A person who has a small knowledge of math is likely to get

almost consistent low score on the test, if the test is given to him twice.

To put it in a nutshell, if a test produces reliable scores, it does not necessarily give us valid results, but if test scores are valid, they must be reliable. Therefore, validity is the most important characteristic of a test score.

Different Types of Validity

Three important types of validity which are reported in articles are **content validity**, **criterion related validity** and **construct validity**. Each will be discussed below.

Content Validity

Content validity refers to the content and format of the measurement instrument. In order to ensure the content validity of a test, you should answer these questions:

How appropriate is the content?

How comprehensive is the content?

Does it logically measure the intended variable?

How adequately does the sample of items or questions in the test represent the content to be assessed?

Is the format of the test appropriate?

Suppose that you have prepared a test of grammar and you claim that your test is a good measure of grammar knowledge. Now you want to ensure that your test has content validity. First of all, it must have adequate sample of grammar rules in English. If you have included only tests that are related to a limited number of grammar rules (only tenses and conditionals are tested) or if you have only included easy or difficult rules, your test will be unrepresentative and you cannot make valid inferences based on the scores. Therefore, the first thing you must do is to make sure that the content is a representative sample of the domain that you want to check.

Second, you must ensure that the format of your instrument is appropriate. By format, I mean that printing must be clear; font size must be big enough; appropriate language should be used and directions must be clear. Ignoring these elements affect the content validity of your instrument. If the test is typed in small fonts and the language of directions cannot be understood properly by the participants in your research, you cannot obtain valid results. Therefore, a competent judge or an expert should look at your test and decide whether it has the above characteristics or not.

Practically speaking, you should take the following steps for the sake of establishing content validity. First, write the definitions and objectives of what you want to measure. For example, you write the objectives of your test: to measure university students' knowledge of grammar, and give the written objectives together with the measurement instrument, and a description of the participants who are going to take the test to one or more experts. The experts go through the questions and according to the objectives of your test, and intended sample tick the questions that need to be replaced or removed and also make comments about the format of the instrument. Then you modify the instrument and give it to the experts for a second review and this process goes on until the experts decide that the test is good in terms of adequacy of sampling and format. There are two important points here: First of all, the experts must have expertise on the subjects of the test and should possess the necessary qualifications for making judgments and second, they should know the characteristics of the intended sample.

Criterion-Related Validity

Sometimes researchers compare scores from the tests or instruments that they have constructed with the scores on an

existing reputable test or an independent criterion. A criterion is a second test or other instrument that measures the same variable. That is why this type of validity is called **criterion-related validity**.

For example, if a new instrument has been constructed to measure general English proficiency, the instrument is given to a sample of students and then the scores of the sample on the instrument are collected. Then the same students are given an existing reputable general English proficiency test such as the TOEFL and their scores are collected. Now the scores of the students on the newly constructed instrument or test are compared with their scores on the TOEFL. If the new test is a valid measure of general English proficiency, its scores must be correlated with the scores on the TOEFL. That is to say, the students with high scores on our new test must have high scores on the TOEFL and the students with low scores on the new test should get low scores on the TOEFL. If this happens, our new test has criterion-related validity. This type of criterion-related validity is called **concurrent validity**. Why do we call it concurrent validity? Concurrent means "at the same time" and since the new test and the criterion test, (the TOEFL in our example) are given to students at the same time

or with only a short time interval, we call this type of validity concurrent validity.

There is another type of validity which is called **predictive validity**. For example, a researcher might administer a new language aptitude test to a group of university students and later compares their scores on the test with their end-of-term grades in language classes. Here the researcher wishes to see to what extent the new aptitude test can predict the final exam score at a language institute.

In both forms of criterion-related validity which are concurrent and predictive validity, a correlation coefficient is obtained. A **correlation coefficient,** which is symbolized by the letter *r*, indicates the degree of relationship that exists between the scores participants or individuals obtain on the measures. If you are already familiar with correlation, you know that a positive relationship shows that a high score on one of the instruments, for example our new test, is accompanied by a high score on the other, for example the TOEFL, and a low score on our new instrument is accompanied by a low score on the criterion test. A negative relationship shows that a high score on one instrument is accompanied by a low score on the other. All correlation coefficients fall between +1.00 and -1.00.

An *r* of .00 indicates that no relationship exists. This correlation coefficient is called **validity coefficient** because it is used to estimate the amount of validity.

Construct Validity

Anxiety, motivation, intelligence, proficiency are constructs because they are the states that exist; they go on inside people. Everyone accepts that they exist, but they cannot be observed directly and researchers have *constructed* these terms (i.e., anxiety, motivation, intelligence) to refer to those states. These constructs must be observed indirectly through tests and measurement instrument. Therefore, if you claim that your test has **construct validity**, you should show that such states are psychologically real or they really exist and your test measures them.

Researchers attempt to collect different types of evidence that will allow them to claim that a test has construct validity. In showing that a test has construct validity, different pieces of evidence are obtained even from content and criterion-related validity. The more and more different pieces of evidence are provided, the better it is. The steps in obtaining construct validity include 1) defining the variable which is being

measured clearly 2) forming hypotheses based on a theory about how people who have much of that construct are different from those that have little of it and 3) testing the hypotheses empirically. In order to show a test has construct validity, the researcher should do one or a combination of three following things:

1-The researcher should provide correlational evidence that a construct has a strong relationship with a certain related variable and a weak relationship with unrelated variables. In fact, construct validity is explored by investigating its relationship with other constructs, both related (convergent validity) and unrelated (discriminant validity).

2- The researcher needs to show that the individuals in one group get higher scores than individuals in another group on the instrument because they possess the construct.

3- The researcher must conduct a factor analysis on scores from the new instrument. For example, anxiety is a theoretical construct and can be seen in a person's behavior. Since anxiety is one construct, it should have one dimension. Therefore, we can conduct a test on all the items in an anxiety questionnaire

to find whether all items in the questionnaire are related to the anxiety construct. The type of test used for this purpose is factor analysis which was explained before.

Let me give you an example here. Suppose that you have designed an English pronunciation test. If you present the following pieces of evidence, you have supported your claim that your test has construct validity.

1- Independent judges all state and show that all items on the pronunciation test require pronunciation ability.

2- Independent judges all agree that the test format, directions, scoring, and reading level would not prevent individuals in answering the tests and revealing their pronunciation ability.

3- Independent judges all agree that the sample of tasks included in the test is relevant and representative of pronunciation tasks.

4- A high correlation exists between learners' scores on the test and scores on the pronunciation part of a standardized test.

5- The students who have had received specific training and teaching in pronunciation have achieved high scores on the test.

5- A high correlation exists between scores on the test and teacher ratings of ability in pronouncing words.

6- Those who are native speakers of English or have studied English as a major obtain higher scores on the test than those who are non-natives and have never studied English.

Let me give you more tangible examples for how you can provide evidence for the construct validity of a test. Suppose that I have a reading test of Persian. I claim that my test measures the construct of reading ability. Therefore, I try to show that my test can differentiate between two groups: a group that poses the construct of reading ability of Persian and a group that lacks it. I find two groups of students who are similar in all ways except that one group is made of nonnative speakers of Persian with very little Persian reading comprehension ability while the other group has a high ability in reading Persian. For example, freshmen of Persian and seniors of Persian at a university in England are selected. I give both groups the test. If the senior students get a high score on

the test and freshmen get a low score, I can argue that my test measures a construct which is called reading ability. Of course, this evidence is not enough. I should accompany other pieces of evidence such as content and criterion related validity to provide evidence that the construct which is measured by the test is reading ability.

In another example, suppose that I have a test which I claim measures the construct of listening comprehension ability. I may choose only one group and give the test to the group as a pretest. Then I go on to train that group in listening comprehension by using listening comprehension materials in class and I give them the same test at the end of the course. If the students get low scores in pretest and high scores at the end of the course, I can claim that my test measures something or a construct and based on the content of what I have practiced throughout the course, that something or construct is listening comprehension.

Important points about validity

Validity is a characteristic of the data and not the instrument. Therefore, before you use an instrument make sure that it produces valid data. The validity of data will be affected

whether it is collected from people who are too young, cannot read or lack motivation. Therefore, if another researcher makes claims of validity, you should first examine the sample of people who took that researcher's test and the conditions under which the test was administered and then use that test for your research purpose before claiming for validity.

When reporting the content validity of your instrument, you must mention who examined the content, what they were asked to do and what their evaluative comments were.

Like reliability, in validity we deal with estimates and not definitive statements. When you want to administer a test to a sample to establish criterion or construct validity, try to have a big enough sample because validity coefficients based on small samples are more likely to change from one situation to another situation.

Unit 12

Internal and External Validity in Research

Internal and External Validity in Research

Internal validity is related to interpreting the findings of research within the study itself. If you can show that the results of your study are only due to treatment, your study has high internal validity. **External validity** is related to generalizing the findings beyond the study and therefore is sometimes called **generalizability**. External validity has to do with whether the findings can be generalized to other people and other contexts. Therefore, if the sample is unique or the conditions, context, and situation are far from the real word, the external validity is under question. Let's see what factors affect the internal and external validity of a research project.

Threats to Internal Validity

The most common type of threats to internal validity can be classified in three broad categories. Some of these threats happen in the process of conducting a study which we refer to as **experience factors**; some of these threats are related to the subjects and participants which we will refer to as **participant factors**; and finally threats which are related to the measurements and instruments which we will refer to them as **instrument factors**.

Experience Factors

History: Suppose you are conducting a study and the participants in one group are suddenly provided with a new extracurricular English class at school and receive more exposure to English. Are your groups still similar? Certainly not.

Testing: Think whether your pretest informs the participants about the topic of your research and whether they studied that topic on their own at home or they discussed the answers to the pretests with their classmates and came up with the answers. Are you sure the results of the posttest are due to your treatment? Certainly not.

Expectancy: Do you expect your treatment group to be better than the control group from the beginning of your research? How about your students? These are called **researcher and subject expectancy**. Can your expectancy affect the results of your study? It might.

Participant Factors

Subject selection: Do the participants whom you selected for your study have different languages, ages, levels of education,

etc. Are you sure there are not any preexisting differences between them?

Maturation: Children's linguistic level increases with age. Are you sure the results of the study are because of treatment and not the increase in linguistic level based on age?

Statistical regression: Learners' scores change over time because of chance factors. High scores at pretests decrease towards the mean and low scores increase towards mean at posttests. Can you do anything about this?

Experimental mortality: Students might drop out of the study. What will you do if some of them in the control or treatment group do not take the posttest or quit? Can you easily exclude them from your sample if you have used one type of random sampling? Are the remaining participants still representative of the population?

Instrument Factors

Reliability: The measurement instruments and those who measure and observe performance might not be consistent across groups and stable across times. Do the raters in your

study act consistently across tests? Do the observers in your research project act consistently across classes?

Threats to External Validity

Keep in mind that if your research project lacks internal validity, it does not have external validity and you cannot generalize from your findings. Therefore, internal validity is a prerequisite for a claim to external validity.

External validity mostly has to do with sampling selection. In fact, the type of sampling that you do can determine whether your study has external validity or not. The best type of sampling that increases external validity is random sampling. If you think that through random sample, you might have more males or females in your sample, or the proportions of specific subjects might be more than the other subjects, you should use stratified random sampling. In this type of sampling you decide ahead of time what portion of the sample should be male/female, educated/uneducated, and then the subjects are randomly selected by category. From this discussion, it can be concluded that if random sampling improves external validity, convenience sampling decreases the external validity. There are other issues besides sampling which affect external

validity. They include **reaction to testing**, **reaction to experiment** and **multiple-treatment interference**.

Reaction to testing: Can part of the improvement be due to using a pretest and alerting the students to the topic? How about the real world in which there is no pretest?

Reaction to experiment: Can the improvement be because of the fact that the participants knew they were under investigation? What if the students studied harder because they knew you were conducting research on them?

Multiple-treatment interference: What if some of the students in the treatment group share the materials in their class with those of the control group?

Unit 13

Statistics: Descriptive Statistics

Descriptive Statistics

In research and statistics, we usually deal with two terms: **population** and **sample**. Population means any group of individuals who are similar in some respects. Therefore, a group will be population when individuals have at least one common characteristic (e.g., the population of university students because they all study at the university).

There are many university students in Iran in different majors who are doing their undergraduate and postgraduate programs. In fact, most of the time the population size is usually big and the people in a population have many characteristics and thus **size** and **diversity** are two problems related to population. To solve the problem of size, we get a **sample** from the population and to solve the problem of diversity, we limit the population to a specific group to which we wish to generalize our findings. The new narrowed down population is called **target population**.

Descriptive statistics simply is used to describe the sample. In the method section of an article, you should report the characteristics of the sample that you selected for the study.

These characteristics include the number of people or cases in your sample, the number or percentage of males and females, the educational level, the mean score or average of the students, standard deviations, the number of the participants, and the minimum/maximum number that the participants got on the tests. This is called **descriptive statistics.**

Moreover, before you conduct some statistical tests such as a t-test, an ANOVA, etc., you should be sure that some assumptions are met. It means that using these techniques depends on certain characteristics of the data that you have. If these characteristics are not in your data, you are not allowed to use the t-test, ANOVA, or certain other statistical techniques. In order to test those assumptions, you need to have descriptive statistics.

You should be careful that descriptive statistics do not allow you to make conclusions from the data and you cannot confirm or reject your hypotheses based on them. Through descriptive statistics, you just describe your data and do not reach any conclusions which are generalizable. As will be explained later, when you wish to make inferences about the population, **inferential statistics** should be used. Therefore, keep in mind

that descriptive statistics and inferential statistics are completely different.

Importance of Descriptive Statistics

Descriptive statistics are very important because if we only present our raw data, it would be difficult to visualize what the data are showing. By raw data, I mean the scores of individual students, for example. Descriptive statistics help us to present the data in a more meaningful way, and as a result, it allows simpler interpretation of the data.

Suppose that we have the scores of 200 students' essays. We may be interested in 1) the overall performance of those students and also 2) the distribution or spread of the scores. Descriptive statistics allow us to do this. Typically, there are two general types of statistics that are used to describe data: One is **measures of central tendency** which allows us to have an idea of the overall performance of those students and the other one is **measures of spread** which enables us to understand the distribution or spread of the scores.

Measures of Central Tendency

Measures of central tendency are ways of describing the central position of a frequency distribution for a group of scores. The frequency distribution refers to the distribution and pattern of scores of, for example, 200 students' essays from the lowest to the highest. We can describe this central position by using a number of statistics, including the **mode**, **median**, and **mean**.

The **mode** refers to the most frequently seen score in the set of scores. Therefore, you look at a set of scores and the score that many students got is the mode. Now if we have set of scores such as 18, 17, 15, 19, 11, 16, 17, 17, we see that 17 is observed more than the other scores. Therefore, 17 is the mode. Sometimes, there is more than one mode in a set of scores. So we call a set of scores with more than one mode **bimodal** and the type of distribution of scores as **bimodal distribution.**

The **median** is the score in the middle of the set of scores. If you want to compute the median, you just list all scores in an order from the lowest to the highest, and then find the score in the center of the set. For example, if our test scores are 18, 17, 15, 19, 11, we arrange them into 11, 15, 17, 18, 19. Then we look at the middle score which is 17. Therefore, 17 is the

median. Remember that if we have even numbers and there is no middle score such as in 11, 15, 17, 18, the median is the average of the middle scores. In this example, the median is 16 (15+ 17 = 32/2= 16).

The **mean** which is also called **average** is the most commonly used method of describing central tendency. In order to calculate the mean, you only add up all the values and divide by the number of values. For example, the mean of a test score, which most people know how to calculate, is computed by summing all the scores and dividing by the number of students taking the exam. For example, if our test scores are 18, 17, 15, 19, 11, the mean is 16.

Measures of Spread

Measures of spread which are sometimes called **measures of dispersion** or **measures of variance** describe how spread out the scores are. For example, the mean score of our 200 students may be 70 out of 100. However, not all students have scored 70. The scores are usually spread out. Some have lower and others have higher scores. Measures of spread help us to summarize how spread out these scores are. To describe this

spread, a number of statistics are available to us, including the **range**, **variance** and **standard deviation**.

The **range** is simply the highest value minus the lowest value. Suppose we have these scores: 18, 17, 15, 19, 11. The highest score in this example is 19 and the lowest score is 11. Therefore, 19-11= 8. Our range is 8. The higher the range, the more spread the distribution is.

Another measure of spread is **variance.** As you know the mean is simply the average of all scores. However, the variance measures the average degree to which each score differs from the mean. The greater the variance, the larger the overall data range. For the variance, first the differences between each score and the mean are calculated. Then, the results are squared and averaged to produce the variance. Considering the above example, if we have scores of 18, 17, 15, 19, 11, the mean will be 16. Then the differences of each score from the mean are squared, (2^2, 1^2, -1^2, 3^2,-5^2), and then added up, ($2^2 + 1^2 + -1^2 + 3^2 + -5^2 = 40$), and are finally divided by the number of scores to get the average, ($40/5 = 8$).Therefore, the variance of these scores is 8. You might ask “Why should we square the

differences?". The answer is that if we just add up the differences from the mean, the negatives cancel the positives.

Another measure of spread which is more accurate than the range is the **standard deviation which is commonly written as SD.** Standard deviation is simply the square root of the variance. **Why do I call it more accurate? Because the range depends on two extreme scores in a class. If there is a very top student and a very weak student in the same class, the range will be a very high score and we might think that the scores are very spread out in the class.** Let's see how standard deviation is calculated. Standard deviation is simply the square root of the variance. After you calculate the variance, you can easily compute standard deviation. For example, if the variance for a set of scores is 8, the square root of 8 is 2.82. Therefore, the standard deviation is 2.82.

An important point to remember is that when you get a sample from the population as you do in most of your research studies, the formula for calculating standard deviation will be a bit different. The only way sample standard deviation formula differs from the population standard deviation formula is that you subtract 1 from the number of participants or scores in the

denominator. That is to say, in the denominator you will have N-1. In fact, you add up the squared differences and divide them by the number of scores minus 1 (i.e., N-1) before you calculate the square root to obtain the standard deviation.

One of the questions that students usually ask is about the difference between variance and standard deviation. The standard deviation is expressed in the same units as the mean, but the variance is expressed in squared units. We use squared numbers for calculating variances to prevent differences below the mean from canceling out those above, which can sometimes result in a variance of zero. Therefore, the variance is no longer in the same unit of measurement as the original data. Taking the root of the variance means the standard deviation is restored to the original unit of measure. You can use either of them as long as you are clear about what you are using. In other words, if you report one of them, you don't need to report the other. For example, a normal distribution with the mean of 10 and the standard deviation of 3 is exactly the same thing as a normal distribution with the mean of 10 and the variance of 9. On the following page, two examples of descriptive statistics are

Sample Descriptive Statistics

Table 1

Descriptive Statistics for the Proficiency Test

Groups	N	Mean	SD	Minimum	Maximum
Experimental Group	19	23.89	6.45	16	42
Control Group	19	24.17	4.58	18	35

As can be seen in Table 1.1, in terms of the obtained mean scores on proficiency test, the two groups averaged 23 to 24 out of possible 50. The analyses revealed that control group performed better than experimental group in terms of obtained mean score. Moreover, as the standard deviations indicate, the control group had lower within-group variability than the experimental group.

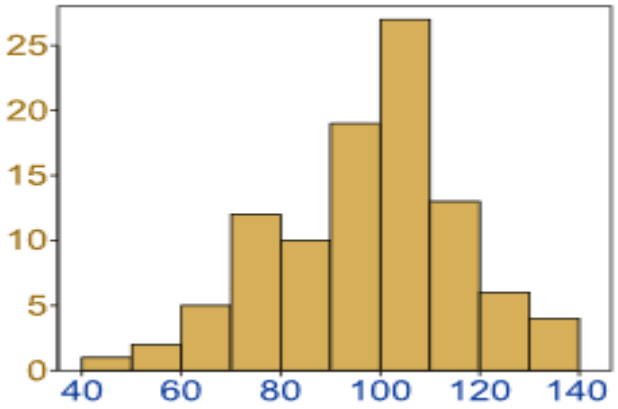
Importance of Graphs

There is a famous saying that "a picture is worth a thousand words." The same is true about a graph. Good graphs convey information quickly and easily to the readers. Graphs can show relationships or differences that are not obvious from studying a list of numbers in tables. Therefore, it's a good idea to supply your descriptive statistics with graphs so that readers can compare or see a trend in the data better. There are 6 main types of graphs which are usually used in articles and theses. They include **histograms, bar graphs, scatter plots, box**

plots, line graphs and **pie charts.** They will be explained below.

Histogram

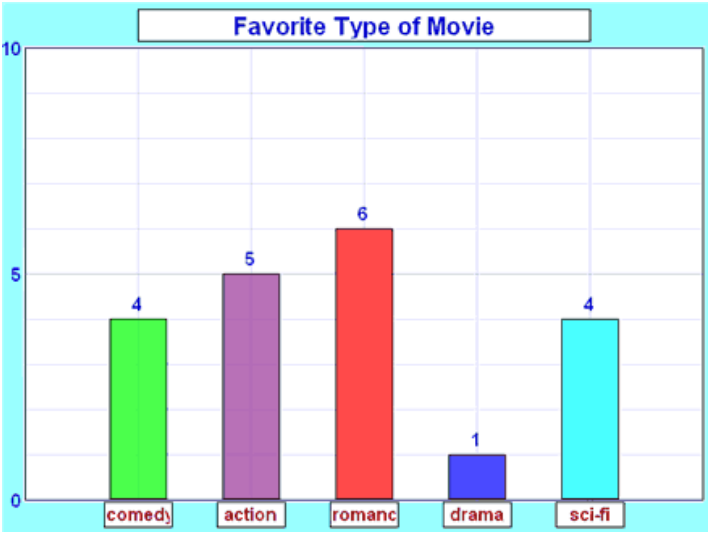
Histograms are used to help researchers better understand how frequently or infrequently certain scores or numbers occur in a given set of data. This type of graph shows whether our data are normally distributed or not. In a histogram, we have just **one continuous variable** (i.e., ages, scores, weights, heights, amount of time). The scores or ages, for example, will be on the horizontal axis and frequencies will be on the vertical axis. For each number or a range of numbers and its frequency, we have bars. A tall bar shows that a score has a high frequency and a short bar shows that a score has a low frequency. An example of a histogram is provided below.



Bar Graph

A **bar graph** which is also called a **bar chart** is similar to a histogram in both in form and purpose. The difference is that in the horizontal axis of a histogram we have a **quantitative variable**, but in the horizontal axis of a bar graph we have **a categorical variable**. Moreover, in a bar graph ordering of the bars can be based on the way that the researcher likes, but in histogram we have a logical order from the lowest score to the highest score. Simply put, in bar graphs we have two variables: one categorical variable and one continuous or quantitative variable. Bar graphs are good when your data are in categories such as gender, groups, etc.

Imagine you just did a survey of your classmates to find which kind of movie they liked best. Here movie is a categorical variable that has different categories such as action, comedy, etc., and the frequency of the classmates who liked different categories is a continuous or quantitative variable. The bar chart below shows the results. As can be seen, romance is the type of movie that most of your classmates like.



Scatter Plot

When you have two continuous variables such as age, height, anxiety, etc. you can show the relationship between them graphically by using a **scatter plot**. A scatter plot has the following uses:

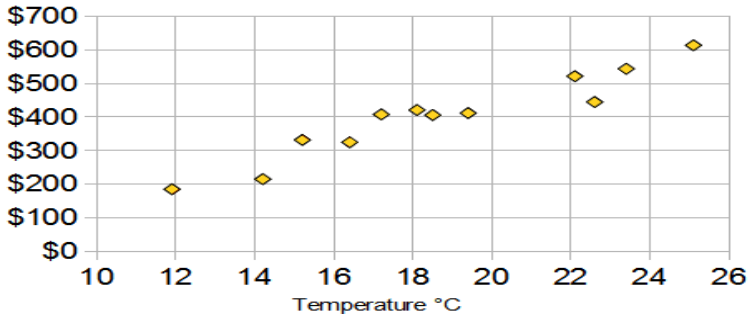
1- It helps you to see if two variables in the study are positively or negatively correlated.

2- It indicates whether the relationship between two variables is linear (i.e., as one variable increases, so does the other variable or as one variable increases, the other decreases) or curvilinear (i.e., a type of relationship between two variables where as one variable increases, so does the other variable, but

only up to a certain point, after which, as one variable continues to increase, the other decreases.

3- It also shows the strength of the relationship.

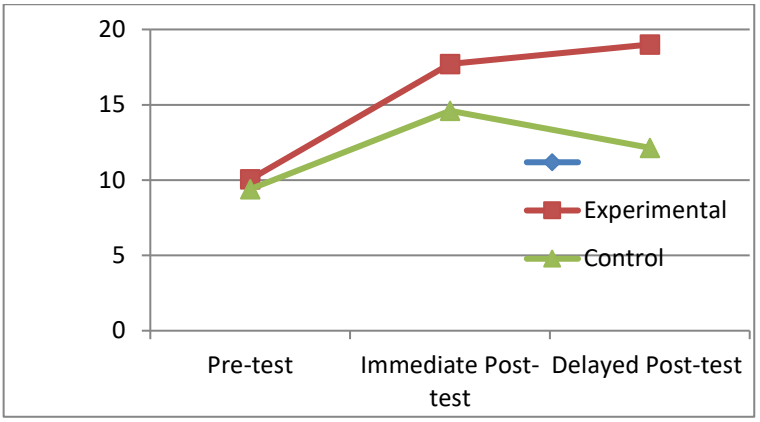
An example of a scatter plot is presented below. It shows the relationship between the temperature and ice cream sales.



As you can see, the relationship between temperature rise and the rate of sale is positive. As the temperature rises, the sales rise too. It also shows that the relationship is linear, but the relationship is not perfect.

Line Graph

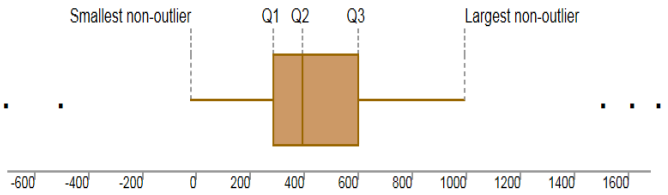
Line graphs are the best types of graphs when you want to compare two or more mean scores across times such as Time 1, Time 2, and Time 3. In a line graph, points are connected by lines to show how something changes in value as time goes by, or as something else happens.



As you can see in the above line graph, the experimental group and the control group had improvement from pretest to posttest, but only the experimental group improved from the immediate posttest to the delayed posttest.

Box Plots

A box plot or **a box and whisker plot** is a very useful tool to compare the distribution of scores on one or more variables separately. It gives the researcher information on outliers, patterns of scores for various groups, the variability in scores within each group and also allows the reader to inspect the differences between the groups visually. A **box plot** consists of a box and two lines extending from the sides of the box which are called **whiskers**. A line which is in the box is called **median**. An example of a box plot is displayed below.

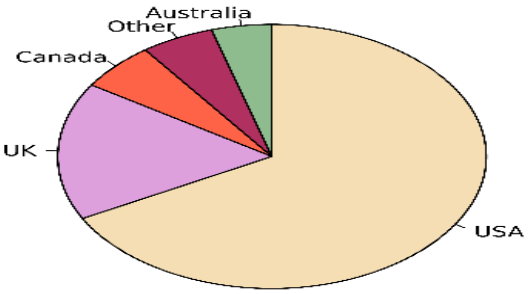


If there are **outliers** (i.e., an extreme value that differs greatly from other values in a set of values) in your data, they are shown separately as points on the chart. In the box plot above, two outliers are before the first whisker and three outliers follow the second whisker. You can also see Q1, Q2, and Q3 in the figure. They are **quartiles.** Quartiles divide a rank-ordered data set into four equal parts. In the above example, the data are ranked from the lowest to the highest first and then are divided into four quartiles. The values that divide each part are called the first, second, and third quartiles and they are denoted by Q1, Q2, and Q3, respectively. In sum, a box plot presents five sample statistics in a visual display: the minimum, the lower quartile, the median, the upper quartile and the maximum. The width of the box (i.e., Quartile 3 minus Quartile 1) is called interquartile range.

Pie Chart

A **pie chart** refers to a type of chart which has the shape of a pie or circle. It shows the relationship of different parts of the

data. Through looking at the pie chart, the reader can easily see the biggest or smallest share of the total data. It is used to show percentage or proportional data and is good for displaying data for around 6 categories or fewer.



Unit 14

Statistics: Inferential Statistics

Introduction

As was mentioned before, in descriptive statistics you just describe what the data are or what the data show, and cannot reach any conclusions which are generalizable. By using **inferential statistics,** you try to reach conclusions that go beyond the immediate data or the data that you have obtained through analyzing the sample. For example, we use inferential statistics to infer from the sample data what the population might think or how the population may behave. We use inferential statistics to make judgments of the probability that a difference that we have observed between the control group and experimental group is not by chance and if we repeat the same study several times, it is very likely that we get the same results. Therefore, you can use inferential statistics to make inferences from your data to more general conditions.

Correlation

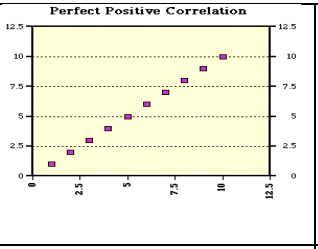
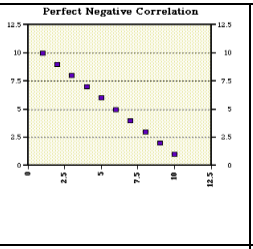
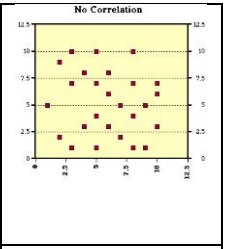
The word *correlation* is made of *co-* (meaning "together"), and *relation*. Sometimes you ask yourself whether the students who are good at math are also good at English. Here you want to know whether these two variables are related. In other words, you want to know whether two sets of data are strongly linked together. If in our example math and English are related, we

can say that those who get high scores on English also get high scores on math (positive relationship) or those who get high scores on English get low scores on math (negative relationship). Therefore, correlation usually has one of the two directions: positive or negative. Correlation is **positive** when the values **increase** together, and correlation is **negative** when one value **decreases**, the other increases. Correlation analysis is sometimes referred to as **bivariate correlation** (*bi* means two and *bivariate correlation* means correlation between two variables). We try to answer questions such as "Is there a relationship between____scores and_____scores?"

In statistics, the correlation coefficient *r* measures the strength and direction of a linear relationship between two variables. The value of r is always between +1 and –1. 1 shows a perfect positive correlation. 0 shows no correlation (the values don't seem linked at all) and -1 shows a perfect negative correlation.

As was explained in Chapter 2, scatter plots will often show whether a relationship exists between two sets of data. Scatter plots usually consist of a large body of data. The closer the data points come when plotted to making a straight line, the higher the correlation between the two variables, or the stronger the

relationship. Scatter plots help show correlation between data. By placing one variable on the X-axis (vertical axis) and the other on the Y-axis (horizontal axis), it is possible to see how changes in one variable change with the other. Scatter plots do not connect dots. Below you can see some examples of scatter plots.

		
Positive correlation	Negative correlation	No correlation

Correlation does not mean causation.

One of the common mistakes that some novice researchers make is that when they see two variables are related, they think that one variable causes the other variable. For example, on hot days people drink fruit juice a lot, and people also go swimming. Therefore, if you calculate the correlation between fruit juice sales and the frequency of going swimming, you find that there is a correlation between fruit juice sales and the number of times people go swimming (they both go up as the temperature goes up in this case). But just because fruit juice

sales increase does not mean fruit juice causes (causation) more swimming or vice versa. Correlation does not always mean that one thing causes the other thing because something else might have caused both, which is hot weather in this case.

Information to look for in a correlation

There are four important pieces of information that you need to consider when reporting correlation. For the sake of simplicity, they are mentioned as steps.

Step 1: Decide whether the correlation coefficient is statistically significant or not. You simply look at the value of significance (sig.). If it is less than .05, you refer to the correlation as statistically significant.

Step 2: In case the correlation is significant, determine the direction of the relationship. Is the relationship between two variables in your study positive or negative? You simply look at the sign + or – before r.

Step 3: What is the size of the value or the correlation? In other words, you need to report how strong the correlation is. For the purpose of this step, look at the value of r. Since

correlation ranges from -1 to 1, the closer the correlation to these numbers, the stronger the relationship.

Step 4: Determine how much variance two variables share. It means how much overlap exists between two variables. The value that shows how much overlap exists between the two variables is called **coefficient of determination**. In order to determine how much variance the two variables in your study share, multiply *r* by itself and then multiply it by 100. For example if r is .6, the coefficient of determination will be .6 × .6 =.36. It means that the two variables share 36 percent of variance. Coefficient of determination can be considered as the effect size (see Chapter 5) for correlation effect size.

Therefore, significance level, the direction, the strength, and the coefficient of determination are the things you should consider when you look at a correlation table.

Sample correlation text

A Pearson's *r* correlation was conducted between motivation scores and scores on a proficiency test. Preliminary analyses revealed no violations of the assumptions of normality,

linearity and homoscedasticity. Motivation showed a strong positive correlation with proficiency, r (90) = *.87, p<.05.*

Partial correlation

Like correlation, **partial correlation** refers to a measure of the strength and direction of a linear relationship between two continuous variables whilst controlling for the effect of other continuous variable which is usually called a ***covariate*** or a **control variable**. We must be careful when we are interpreting correlational data. If two variables, A and B, are correlated, it is difficult to say why such a relationship exists. Let me give you an example. Suppose that a study which was conducted on the children between 2 years and 10 years of age showed that the children who have big fingers know their mother tongue better than those who have small fingers. Now a question needs to be answered? Is it really so? Can we say that whenever we see a person with a big finger, he/she knows his/her mother tongue very well? Well, here you should consider the presence of another variable: age. As the age of children increases, their fingers also get bigger and their language skills get better. In fact, age here is a **mediating,** or **confounding variable**. It is a confounding variable because it is related to both finger size and language skill and therefore it explains the relationship between finger size and language skill. In this example if we

control for age, the correlation of language skill and finger size disappears. In cases such as these, when we doubt that the relationship between two variables might be explained by another variable, we should use a partial correlation.

Thus, partial correlation tries to answer questions such as “What is the relationship between English test scores and math test scores after controlling for hours that the students have spent studying?” or “After controlling for age, what is the relationship between intelligence and motivation?”

We conduct partial correlation when the third variable has shown a relationship to one or both of the primary variables. In other words, we should first do a correlational analysis on all variables so that we can see whether there are significant relationships amongst the variables, including any "third variables" that may have a significant relationship to the variables under investigation. Moreover, you need to have some theoretical reasons why the third variable might affect the relationship between the two variables under investigation before you decide to perform a partial correlation.

An important point to remember is that the conclusions that we make in scientific studies are not definite and certain and we should be careful when we interpret correlational data and make conclusions. For instance, even in the finger size and language skill example, we cannot for sure claim that changes in age cause changes in the size of finger or changes in language skills. We can only argue that age **accounts for** or **explains** the relationship between finger size and language skill. Why? Because there are other mediating variables which we might not know and might be explored in the future by other researchers or even us and can result in modifying our interpretation.

Sample partial correlation text

A partial correlation controlling for age found a strong negative correlation between length of residence and production accuracy of English words. The Pearson r statistic was negative (r= -.75, p=.002), meaning scores on production accuracy decreased with increasing length of residence, and the effect size was large (R^2 = .56). Controlling for age, no correlations were found between length of residence and scores on the language aptitude test (r = .03, p = .93).

Regression

Regression analysis is used when you want to predict a continuous dependent variable (i.e., a dependent variable that can take on any value between two specified values such as height) from a number of independent variables. However, the independent variables used in regression can be either **continuous** (variables that can take any number such as height) or categorical (i.e., variables that have two or more levels such as gender). If the dependent variable is dichotomous, then **logistic regression** should be used.

Regression is like correlation because it is concerned with relationships among variables. But correlation and regression differ in three respects: **their purpose**, **the way variables are labeled**, and **the kinds of inferential tests which are applied to the data.**

The first difference between correlation and regression is about the purpose of each technique. Correlation is designed to reveal the relationship between two variables and the correlation coefficient may indicate that the relationship is positive and strong, or negative and moderate, or weak. In correlation each of the two variables is equally responsible for the nature and

strength of the link between the two variables. However, regression tries to make predictions or provide explanations based on one of the variables in either end of the link. For example, a researcher might wish to know how he could predict a student's university GPA (i.e., Grade Point Average) if he knows his or her high school GPA.

Sometimes regression is used to explain why certain people score differently on a particular variable. For example, a researcher might be interested in why university students differ in the degree to which they seem happy with the quality of education at their university. For the purpose of the study, a questionnaire is administered to a large group of university students for the purpose of measuring their satisfaction with the quality of university education. Those same students are also measured on several other variables that might explain why some students are content with the quality of education whereas other students complain nonstop about the low quality. Such variables might include previous university experience, level of parents' education, and gender. Then after conducting regression, the researcher may argue that previous university experience and gender explain the amount of satisfaction. In other words, the extent to which a person is satisfied with the

quality of education at the university depends on previous university experience and gender.

The second difference between correlation and regression is that in correlation, the variables do not have any names such as independent or dependent variables, but in regression variables have labels such as dependent and independent variables. In fact, it is necessary for one variable in regression to be independent variable which is sometimes called **predictor variable** and the other one to be the dependent variable which is sometimes called **criterion variable.**

The third difference between correlation and regression is that in correlation, there is just one thing that we usually focus on and that thing is the sample correlation coefficient. But in regression, we focus on the correlation coefficient, the regression coefficient, the intercept, the change in the regression coefficient, and something called the odds ratio. I will not go into details about these here.

There are different types of regression techniques which can help researchers make predictions. These techniques differ based on **the number of independent variables**, **type of**

dependent variables and **the shape of regression line**. I will discuss the three frequently used techniques here: 1) **bivariate regression**, 2) **multiple regression**, and 3) **logistic regression**. Bivariate regression is similar to bivariate correlation, because both are designed for situations in which there are just two variables. Multiple regression, is created for cases in which there are three or more independent variables. Logistic regression is used for cases in which our dependent variable is dichotomous.

Bivariate regression

This type of regression is very simple. In this type of regression, we have only two variables. One of the variables is the dependent variable and the other is the independent variable. This type of regression can be used for either prediction or explanation. Through this type of regression, the researcher tries to see how well scores on the dependent variable can be predicted from data on the independent variable.

Sometimes in real lives we ask others to predict something for us. For example, we ask our doctor to predict how long it takes us to lose 10 kg of weight. Normally, we expect the doctor to

answer this question by considering our individual case. If the doctor has access to information about other people, he can use regression to make predictions.

Multiple regression

In **multiple regression**, we have more than one independent variable. In other words, there are two or more independent variables. Similar to bivariate regression, the research uses multiple regression for either prediction (i.e., predicting the dependent variable) or explanation (i.e., explaining the independent variables which cause something). The main difference between multiple regression and bivariate regression apart from the number of independent variables is that in multiple regression we can control one of the independent variables as *covariate* and also examine the *interactions* between independent variables. There are three main types of multiple regression: 1) **simultaneous multiple regression** 2) **stepwise multiple regression**, 3) **hierarchical multiple regression.**

Simultaneous multiple regression

In **simultaneous multiple regression**, which is sometimes called **standard regression** and is the most commonly used type of multiple regression analysis, the data related to all

independent variables are considered at the same time. In other words, in standard multiple regression, all the independent variables are entered into the equation at the same time. You should use this type of regression if you have a number of variables such as motivation, aptitude, intelligence and want to know how much variance in a dependent variable, such as university academic success, these variables can explain as a group or block or separately.

Stepwise multiple regression

In **stepwise multiple regression**, it is the computer which determines the order in which the independent variables should become a part of the regression equation. In this type of regression, the researcher provides the computer and software with a number of independent variables and then waits for the program to select which variables and in which order those variables should enter the regression equation.

Hierarchical multiple regression

In **hierarchical regression** which is also known as **sequential regression**, the researcher himself/herself decides which independent variables and in which order they should be entered into the equation. He usually does this based on theoretical grounds. The researcher tries to find how much or

how well each independent variable contributes to the prediction of the dependent variable after other variables have been controlled for. For example, if the researcher wishes to know how well motivation predicts success, after the effect of intelligence is controlled for, he enters intelligence in Block 1 and then motivation in Block 2. When all the relevant variables are entered into the software, the overall model and the relative contribution of each block of variables is assessed to see how well the variables or combination of them can predict the dependent variable.

Logistic Regression

Logistic regression is used to predict a discrete or categorical outcome based on variables which might be discrete or continuous. Suppose you wish to predict the absenteeism patterns of teachers based on their individual differences such as experience, knowledge, motivation. Absenteeism here means whether teachers are usually present or absent at work. It is a discrete variable because teachers can be either present or absent and cannot be both. Therefore, the dependent variable is a discrete variable. However, the independent variable can be either discrete or continuous. For example, gender of teachers can be a categorical variable while their experience which

might range from 1 to 50 years of experience can be considered as a continuous variable.

There are a number of points you need to consider before you conduct logistic regression. First, in order for logistic regression to give you good results, you should have a sample that is large enough and also the number of independent variables should match the sample size. If you have a small sample and a lot of independent variables, you cannot be hopeful to get valid results.

Second, the independent variables should not be related to each other. This is called **Multicollinearity**. If your independent variables are highly correlated, you cannot depend on the results. It should be mentioned that correlation between the independent variable and the dependent variable is highly desired, but the correlation between the independent variables is undesirable.

Sample regression text (Logistic)

Direct logistic regression was conducted to assess the impact of several factors on the likelihood that teachers would report that they loved teaching. The model contained 3 independent

variables (gender, attitude, and the amount of salary categorized as high and low). The full model containing all predictors was statistically significant, $\chi 2$ (3, N = 300) = 80.02, $p < .05$, indicating that the model was able to distinguish between teachers who reported they loved their jobs and those who did not report. The model as a whole explained between 30.2% (Cox and Snell R square) and 45.3% (Nagelkerke R squared) of the variance in love for job, and correctly classified 80.2% of cases. As displayed in Table

1, only two of the independent variables made a unique statistically significant contribution to the model (attitude and amount of salary). The strongest predictor of reporting a love for job was attitude, recording an odds ratio of 7.27. This indicated that teachers who had a positive attitude to their jobs were over 7 times more likely to report a love for their jobs than those who did not have positive attitudes controlling for all other factors in the model.

Factor analysis

Factor analysis is a statistical technique that is used to reduce and summarize data. By reducing data, I mean that you have a large number of questions in a questionnaire and you wish to see which questions are related to each other and measure a

specific variable or construct. This specific variable that you try to discover or explore is called a **factor** or a **component** and therefore this technique is called factor analysis. Factor analysis is necessary for developing tests, scales, and questionnaires. For example, when you want to design a questionnaire about class anxiety, you should collect as many items as possible about it. How do you do that? You interview teachers and ask them about the things that they think can cause class anxiety. Then you write some items based on the teachers' data which were collected qualitatively and put all these items in a questionnaire and then pilot the questionnaire. Suppose there are 100 items. After you collect learners' responses on those 100 questions, you enter the data into SPSS and conduct factor analysis to explore and discover which questions are related to a specific factor. This technique is called **exploratory factor analysis** and is used in early stages of research.

There is another type of factor analysis which is called **confirmatory factor analysis**. Confirmatory factor analysis is used to test how well the measured variables represent the number of constructs. You know that constructs refer to those variables that are not observable such as intelligence, aptitude

and motivation. In a confirmatory factor analysis, based on a theory or the findings from previous research, the researcher specifies the desired number of factors and how measured variables are related to those factors. Here, contrary to exploratory factor analysis, the researcher does not intend to explore and discover the factors or questions related to those factors. Rather, the researcher has some idea of the factors and the variables that form that factor, and tries to confirm that those variables are related to the factor. In confirmatory factor analysis, we try to answer questions such as "From my 40-item instrument, are the 4 factors clearly identifiable constructs as measured by the 10 questions that they are comprised of?" or the researcher says "I have a 30-item questionnaire which measures anxiety and do all these 30 survey questions accurately measure one factor (i.e., anxiety)?"

Sample factor analysis text

The 30 items of the Anxiety Scale were subjected to principal components analysis using SPSS version 21. Before performing principal components analysis, data were assessed to ensure its suitability for factor analysis. Examining the correlation matrix showed the presence of many coefficients of .3 and above. The Kaiser- Meyer-Olkin value was .74, which

exceeded the recommended value of .6 and Bartlett's Test of Sphericity reached statistical significance, supporting the factorability of the correlation matrix. Principal components analysis revealed the presence of four components with eigenvalues exceeding 1, explaining 40.1%, 12%, 8.3% of the variance respectively. Based on Catell's (1966) scree test, it was decided to retain two components for further investigation. The results of Parallel Analysis also showed only two components with eigenvalues exceeding the corresponding criterion values for a randomly generated data matrix of the same size (10 variables × 300 participants). The two-component solution explained a total of 39.4% of the variance, with Component 1 contributing 35.40% and Component 2 contributing 26.0%. To aid in the interpretation of these two components, oblimin rotation was run. The rotated solution revealed the presence of simple structure, with both components showing a number of strong loadings and all variables loading substantially on only one component. The interpretation of the two components was consistent with previous studies on the Anxiety Scale, with positive affect items loading strongly on Component 1 and negative affect items loading strongly on Component 2. There was a weak negative correlation between the two factors (r = –.14).

T-test

A **t-test** is used to show whether two mean scores are significantly different. In other words, a t-test tells us if the difference between the two mean scores is large enough to let us say that they are significantly different. A t-test is also used to let us decide whether the mean score of a group at Time 1 is different from the mean score of the same group at Time 2.

Be careful that a t-test is used for the difference between the mean scores of two groups or two occasions. For more than two groups or occasions, you should use another type of statistical test which will be discussed in the next chapter. There are usually three types of t-tests: **one sample t-test, paired samples t-test**, and **independent samples t-test**. In this section, the independent samples t-test and paired sample t-test, which are commonly used in theses and articles, are explained.

Independent samples t-test

Suppose that you have decided to conduct a study and compare the effect of using pictures on vocabulary learning. You select some participants randomly from the population and assign them into two groups randomly. Here you have two groups: an experimental group and a control group. Because you want to investigate the effect of pictures on vocabulary learning, first of all you need to show that the two groups are equal with regard

to vocabulary knowledge. In fact, the two groups should be pretested on vocabulary knowledge first. You give the two groups a test of vocabulary first. The mean score of the participants on the vocabulary test in the experimental group is 17 and the control group is 16. Apparently 17 and 16 are different. But are they statistically different? Here, you need to conduct an **independent samples t-test** to find whether the two mean scores are statistically significant. Why t-test? Because you have two groups. Why independent sample t-test? Because you have two groups that are independent from each other and the participants in each group are not related to participants in another group. What is the purpose of using a t-test? To find whether there is a statistically significant difference between the two groups.

Sample independent samples t-test

An independent-samples t-test was conducted to compare the self-esteem scores for males and females. There was not a significant difference in scores for males (M = 34.02, SD = 4.91) and females (M = 33.17, SD = 5.71) t (434) = 1.62, p = .11, two-tailed). The magnitude of the differences in the means (mean difference = .85, 95% CI: –1.80 to 1.87) was very small (eta squared = .006).

Paired samples t-test

A paired samples t-test which is also called a **dependent samples t-test** and **repeated measures** is used when we want to examine the difference between two mean scores from one single group. Imagine that you have a class and you wish to see whether their motivation increases from the beginning of the term to the end of the term. To investigate this issue, you give the students a motivation questionnaire in the first session of the course and at the end of the course you give them the same questionnaire. Suppose that the mean score of the class is 30 at the beginning of the course and the mean score of their motivation increases to 50 at the end of the course. You want to see whether motivation increased from Time 1 to Time 2. Here, you need to conduct a paired samples t-test to find whether the groups' mean scores at Time 1 and Time 2 are statistically significant. Why paired samples t-test? Because you have one group at two different times. What is the purpose of using a paired samples t-test? To find whether there is a statistically significant difference between two testing times for one group. In the above example, we talked about two times for one single group. A paired samples t-test can also be used when we give the same group two different types of tests. Suppose that you want to see whether students in your class

(single class) are better at grammar or vocabulary. You give a grammar test to your students and also a vocabulary test. Then you score the tests. Here we have one group and two different tests. Therefore, we should use the paired samples t-test.

Therefore, keep in mind that for one group at two different times or one group for two different tests, we should use a paired samples t-test. Note that you can use a paired samples t-test for one person at two different times or one person for two different tests or questions.

Sample dependent samples t-test

A paired samples t-test was conducted to evaluate the impact of the intervention on students' scores on an anxiety questionnaire. There was a statistically significant decrease in anxiety scores from Time 1 (M = 39.18, SD = 4.11) to Time 2 (M = 36.3, SD = 5.15), t (29) = 5.39, p <. 0005 (two-tailed). The mean decrease in anxiety scores was 2.27 with a 95% confidence interval ranging from 1.66 to 3.68. The eta squared statistic (.50) indicated a large effect size.

One-way ANOVA

The one-way analysis of variance which is usually written as **ANOVA** is a test which is used to determine whether there are any significant differences between the means of three or more groups. One-Way ANOVA can also be used for two groups, but statisticians prefer to use an independent samples t-test for two groups.

Similar to t-tests, there are two types of one-way ANOVAs: repeated measures ANOVA which compares the mean scores of one group or person at three or more times and between-groups ANOVA, also known as independent samples ANOVA, which compares the mean scores of three or more different groups.

One-way between-groups ANOVA

As was mentioned before, the one-way between-groups ANOVA is used to determine whether there are any significant differences between the means of two or more independent or unrelated groups. There are two important points that you need to remember. One important point is that the one-way ANOVA is an *omnibus* test statistic. It means that this test cannot tell you which specific groups are significantly different from each other. It only shows you that three or more groups are different.

If you wish to determine which specific groups differ from each other, you should use a *post hoc* test after you perform an ANOVA.

Another important point is that you should conduct a post hoc test only after you find a significant difference between the groups. In other words, if you perform an ANOVA and you see that the significant value is less than .05, you should continue and conduct a post hoc test. In case you see that the results are not significant, you do not need to conduct a post hoc test. Not achieving a statistically significant result does not mean that there is no need to report group means and standard deviations

One-way ANOVA focuses on one independent variable or factor which has three or more levels. Note that words *factor* and *independent variable* are usually used synonymously in statistics. For example if a study investigates the effect of three types of methods such as A, B, and C on language learning, method is an independent variable or factor which has got three levels: A, B, C. Unfortunately some students confuse levels with the number of independent variables. Some researchers use the word ANOVA instead of one-way ANOVA.

Assumptions of ANOVA

There are some assumptions that need to be considered when conducting an ANOVA test. One of the assumptions is **normality.** It means that your data must have normal distribution. Interestingly, the one-way ANOVA is a robust test against the normality assumption on condition that the group sizes are not very different. This means that ANOVA tolerates violations to its normality assumption almost well and can tolerate the data that are non-normal with only a small effect on the **Type I error** rate in situations where group sizes are almost equal. Type 1 error, as mentioned before, means rejecting the null hypothesis by mistake. In case your data are not normal, and you want to be cautious in conducting a good analysis, you had better use the non-parametric **Kruskal-Wallis H Test** which does not require the assumption of normality.

Another assumption is **homogeneity of variance**. There are two options that you can take if this assumption is not met. You can use one of the two tests of (1) **Welch** or (2) **Brown and Forsythe test**. For most situations it has been shown that the Welch test is best. Both the Welch and Brown and Forsythe tests are available in SPSS Statistics.

If you don't want to use any of these testes, you could run a **Kruskal-Wallis H Test**. Keep in mind that if the homogeneity of variances assumption is not supported, you should add into your results section that this assumption was violated and you needed to run a Welch F test or a Kruskal-Wallis H Test.

Post hoc tests

As said before, after the results of your ANOVA test showed that there was a significant difference between the groups because the significant value in the ANOVA table is less than .05 or any other alpha level that you have set for your study, you need to conduct a post hoc test to see which groups are different from each other statistically.

There are a large number of different post hoc tests that you can use after running ANOVA. However, you should only run one post hoc test and should not perform multiple post hoc tests. If your data meet the assumption of homogeneity of variances, either use the **Tukey's honestly significant difference (HSD)** or **Scheffé post hoc tests**. Statisticians recommend **Tukey's HSD** test because it is less conservative than the Scheffé test. It means that you are more likely to find

differences if they exist when you use Tukey's HSD test than when you use the Scheffé test. In SPSS Statistics, Tukey's HSD test is referred to as "Tukey" in the post hoc multiple comparisons dialogue box.

If your data did not meet the homogeneity of variances assumption, you should perform the **Games Howell** or **Dunnett's C** post hoc test although the Games Howell test is mostly recommended.

Sample one-way between groups ANOVA

A one-way between-groups analysis of variance was conducted to examine the effect of age on levels of vocabulary knowledge. Participants were divided into three groups according to their age (Group 1: 29 years or less; Group 2: 30 to 44 years; Group 3: 45 years and above). There was a statistically significant difference at the $p < .05$ level in vocabulary scores for the three age groups: F (2, 432) = 4.6, p =.01. Despite reaching statistical significance, the actual difference in mean scores between the groups was quite small. The effect size, calculated using eta squared, was .02. Post-hoc comparisons using the Tukey HSD test showed that the mean score for Group 1 (M = 21.36, SD = 4.55) was significantly

different from Group 3 (M = 22.96, SD = 4.49). Group 2 (M = 22.10, SD = 4.15) did not differ significantly from either Group 1 or 3.

One-way repeated measures ANOVA

One-way repeated measures ANOVA is similar to a paired samples t-test with the difference that a t-test is only used for two times but one-way repeated measures ANOVA can be used for more than two times. One-way repeated measures ANOVA is used to compare three or more means of the same group. This usually occurs in two situations: (1) when the same group is measured multiple times (pretest, posttest, delayed posttest) to see changes in the group over time or (2) when one group takes several different tests and we wish to compare the same group on several measures. In order to help you understand the use of repeated measures ANOVA, I will give you two examples here. Suppose that you have one class and you wish to see the effect of your teaching on the motivation of students in that class. In this example, your treatment is your independent variable and motivation is your dependent variable. You give the students in your class a motivation questionnaire at the beginning of the term, and then the same questionnaire at the end of the term and again you

give the same questionnaire after one year. Here you should use a repeated measures ANOVA. Why? Because you have one class which takes a test or completes a questionnaire three times: before the treatment, after the treatment and one year later.

Sample one-way repeated measures ANOVA

A one-way repeated measures ANOVA was performed in order to compare scores on the motivation questionnaire at Time 1 (prior to the intervention), Time 2 (following the intervention) and Time 3 (three-month follow-up). There was a significant effect for time, Wilks' Lambda = .43, F (2, 21) = 27.13, $p <$ .05, multivariate partial eta squared = .55.

Using multiple t-tests instead of ANOVA

Sometimes, some university students conduct multiple t-tests instead of one ANOVA. It means that instead of running one ANOVA and comparing the mean scores of group one and two and three, they run a t-test on the means of Group A and Group B, Group A and Group C, Group B and Group C. You should remember that every time you run a t-test, there is a chance that you will make a Type 1 error. Type 1 error in statistics means rejecting the null hypothesis by mistake. It means that if you

run several t-tests (either independent sample or dependent sample t-test), you might find a significant difference between the groups and reject the null hypothesis that states there is not a significant difference between the groups. But if you run an ANOVA, you may not find a significant difference. Therefore, if you have three groups and more, use ANOVA instead of multiple t-tests if you do not want to reject the null hypothesis mistakenly.

Two-way ANOVA

The two-way ANOVA has several names. For example, because a factor is an independent variable, we can call it a two-way factorial design or a two-factor ANOVA. Sometimes researchers label this design with regard to the number of levels of each independent variable or factor. For example, if a study had two levels of the first independent variable and 3 levels of the second independent variable, this would be referred to as a 2 × 3 factorial or a 2 × 3 ANOVA.

The main purpose of a two-way ANOVA is to understand if there is an interaction between the two independent variables on the dependent variable. The difference between one-way ANOVA and two-way ANOVA is in the number of

independent variables. In one-way ANOVA we have one independent variable and wish to investigate its effect on our dependent variable. But in two-way ANOVA we have two independent variables and wish to investigate their effects on our dependent variable at the same time. Let me give you an example.

Suppose that you have decided to examine the effect of three types of methods (A, B and C) on grammar learning. So you have three groups. What is your independent variable? It is the method. How many levels does it have? Three levels. What is the dependent variable? Grammar learning. What statistical technique should you use to compare students' mean scores on post-tests? ANOVA. Why? Because you have one independent variable with three levels and one dependent variable.

Well, now imagine that you believe the independent variable which is method can be more or less effective with regard to gender. In fact, you want to see if your teaching methods are equally effective for both males and females. Therefore, another independent variable is important to you and that independent variable is gender. So how many independent variables do you have now? Two: one is method and the other

one is gender. What statistical technique should you use? Two-way ANOVA. Why two-way ANOVA? Because you have two independent variables. Therefore, keep in mind that "way" in one–way ANOVA or two-way ANOVA refers to the number of independent variables. Again you should be careful not to confuse independent variables or factors with the levels of the independent variable.

Using multiple one-way ANOVAs instead of two- way ANOVA

The main reason that researchers prefer to use a single two-way ANOVA instead of multiple one-way ANOVAs is the same as the reason for carrying out a one-way ANOVA rather than using multiple t-tests. You should remember that if you run several one-way ANOVAs (either independent samples ANOVA or repeated measures ANOVA), you take the risk of committing Type 1 error which means rejecting the null hypothesis by mistake.

However, there is one more reason for carrying out a two-way ANOVA instead of using multiple one-way ANOVA tests. The reason is that carrying out multiple one-way ANOVAs would

not allow the researcher to test for any interactions between the independent variables in addition to the main effects.

Main effect refers to the effects of the individual independent variables on the dependent variable. **Interaction effect** refers to the effect of two independent variables on the dependent variable simultaneously. For example, I wish to know if using music has any effect on vocabulary learning. This is the main effect I can investigate. I also wish to see if the effect of music on vocabulary learning is different for children and adults. This is an interaction effect. Interaction effect shows that for example music can have effect on vocabulary learning but the effect is only there for children and not adults. Therefore, the researcher concludes that there is an interaction effect between two independent variables: using music and age. Therefore, through a two-way ANOVA, we can study the effects of the individual independent variables, called the main effects, as well as the interaction effect which shows how different levels of the two independent variables (i.e., the first independent variable with two levels of music and lack of music and the second independent variable with two levels of children and adults) affect the dependent variable. If the effect is not the

same, we say there is an interaction between the two independent variables.

Sample two-way ANOVA text

A two-way between-groups analysis of variance was conducted to explore the impact of gender and age on levels of motivation, as measured by a motivation questionnaire. Participants were divided into three groups according to their age (Group 1: 18–29 years; Group 2: 30–44 years; Group 3: 45 years and above). The interaction effect between gender and age group was not statistically significant, F (2, 350) = 2.25, p = .12. There was a statistically significant main effect for age, F (2, 350) = 3.80, p = .04; however, the effect size was small (partial eta squared =.03). Post-hoc comparisons using the Tukey HSD test indicated that the mean score for the 18–29 years age group (M = 21.36, SD = 4.55) was significantly different from the 45 + age group (M = 22.96, SD = 4.49). The 30–44 years age group (M = 22.10, SD = 4.15) did not differ significantly from either of the other groups. The main effect for gender, F (1, 350) = .27, p = .44, did not reach statistical significance.

Mixed- between–within ANOVA

Mixed- between–within ANOVA is very much similar to two way-ANOVA because both analysis methods need two independent variables. Mixed between-within subjects ANOVA which is also called **split-plot ANOVA** mixes two different types of one-way ANOVA into one study: *between-groups ANOVA* and *within-subjects ANOVA*. In a mixed-between–within ANOVA, one of the independent variables is time and the other one is group, method, etc. Suppose that you want to see whether teaching vocabulary through pictures results in more learning than teaching vocabulary through music in one month. In fact, there is a very important question here that you want to answer: "Can these two techniques have differential effects from pretest to posttest which is going to be given to students one month after the pretest?" So you say from the pretest to the posttest and this means improvement over time is your purpose of study. Therefore, you want to see which group will perform better in the course of time. This is called interaction effect in statistics. Interaction refers to the way in which a category of one independent variable (method of teaching vocabulary) combines with a category of the other independent variable (time here) to produce an effect on the dependent variable.

Sample mixed- between–within ANOVA

A mixed between-within ANOVA was performed to assess the effect of two different interventions (teaching through pictures and teaching through vocabulary lists) on learners' scores on the Vocabulary Test, across three time periods (pre-intervention, post-intervention and two-month follow-up). There was no significant interaction between method type and time, Wilks' Lambda = .20, F (2, 35) = 1.44, p = .09, partial eta squared = .11. There was a substantial main effect for time, Wilks' Lambda = .44, F (2, 35) = 21.12, p < .05, partial eta squared = .45, with both groups showing an improvement in Vocabulary Test scores across the three time periods (see Table 2). The main effect comparing the two types of intervention was not significant, F (1, 35) = .042, p = .12, partial eta squared = .05, which suggests that there is no significant difference in the effectiveness of the two teaching approaches.

Chi-square goodness of fit test

Chi-square goodness of fit test which is sometimes called **one-sample chi-square test** is applied when you have one categorical variable from a single population. It is used to determine whether sample data are consistent with a

hypothesized distribution. It should be emphasized that in **chi-square goodness of fit test,** we have only one categorical variable. Variables can be classified as categorical or numerical. In order to avoid confusion, I only explain categorical variables here.

Categorical variables are also known as qualitative variables or discrete variables. These variables take on values that are names or labels. The gender of people (e.g., male or female) or marital status of a person (e.g., single or married) would be examples of categorical variables. People are in one of these categories and cannot be all of them. For example a person is either male or female, single or married.

Suppose that an importer of foreign cars who is in the business of importing cars wishes to see whether costumers have any preferences among three makes of cars: Chinese cars, Japanese cars, and Korean cars. He should use a chi-square goodness of fit test here to compare frequencies of people who prefer these cars. Let me explain why. How many variables do we have here? We have one variable. What type is it? It is categorical. What are different levels or categories of the variable? Chinese cars, Japanese cars, and Korean cars. What type of data do we

deal with? Frequencies. Therefore, the importer gets a sample of, for example, 500 hundred people randomly and asks them about their preference for the types of cars. Then the frequency of people who prefer Chinese cars is compared to the frequency of people who prefer Japanese cars, and the frequency of people who prefer Korean cars through a Chi-square test for goodness of fit. If the significance level is less than .05, it shows that there is a significant difference between the frequencies and one or two types of cars are preferred more than other types.

Sample chi-square goodness of fit test.

A Chi-square goodness of fit test was performed to determine whether the three types of strategies were equally preferred. Preference for the three types of strategies was not equally distributed in the population, X^2 (2, N = 55) = 4.53, $p < .05$.

Chi-square test for independence

The **chi-square test for independence** is also known as **Pearson's chi-square test** or the **chi-square test of association**. This test is used to find if there is a relationship between two categorical variables. What needs to be

emphasized here is that in the chi-square test for independence, we have two independent variables which are categorical. The chi-Square statistic compares the frequencies or counts of categorical responses between two (or more) independent groups. For example, we have two categorical variables such as job and a question which can be answered either yes or no. For example a question such as "Do you enjoy studying English?" In the job variable we have two levels: teacher and student. In the question variable, we have two levels: Yes/No. Now we want to see if these two variables are related or not. In other words, we wish to find out if teachers are more likely to answer "yes" or "no" to the question than students. In examples such as these, we use a chi-square test of independence.

Sample chi-Square test for independence

A chi-square test for independence was performed to examine the relation between gender and college interest. The relation between these variables was significant, X^2 (2, N = 120) = 12.16, p <.01. Male students were less likely to show an interest in attending college than were female students.

ANCOVA

In ANOVAs, there are just two kinds of variables: independent variables and dependent variables. The data which are analyzed in ANOVAs are considered as the dependent variable; and one or two factors which are also called independent variables. In fact, ANOVAs can involve more than one factor or independent variables which can be between-subjects or within-subjects factors and factors can also have different numbers of levels. For example, when studying the effect of a teacher's voice on vocabulary learning, your independent variable is teacher's voice which can have different levels such as harsh and soft voice. This independent variable is between-subjects variable because you wish to compare two groups who are exposed to teacher's voice differently and you wish to see if there is a difference between the groups at posttests. Your dependent variable is vocabulary learning scores. Your within-subject variable is each single group's improvement from pretest to posttest. It is called within-subject because you wish examine each group separately and see if they improve from Time 1 to Time 2. Therefore, ANOVAs contain two important elements: one or more independent variables and one dependent variable.

Now let's see what an **ANCOVA** is. In an ANCOVA, which stands for *Analysis of Covariance*, we have three rather than two kinds of variables. Similar to the ANOVAs, there will be scores that are the dependent variable and one or more independent variable(s). In ANCOVAs we have a third variable which is called a covariate variable. Since the covariate is a variable on which the study's participants are measured, it is more similar to the study's dependent variable than to the independent variables, but you should not forget that, the covariate and dependent variables have very different functions in a study in which ANCOVA is used.

ANCOVA is more useful than ANOVA sometimes. ANCOVA is used when groups' pretest mean scores are different. Suppose you have three groups and their mean scores at pretest are 17, 14, 13. You conduct an ANOVA and you find that these mean scores are significantly different. So what should you do? Don't worry. ANCOVA is there. You conduct your experiment and when your treatment is over, you give the groups a post-test. Then conduct an ANCOVA on the posttest. ANCOVA compares the groups on posttests and considers their scores at pretest as a covariate.

In another situation, imagine we wished to investigate how well male and female students at different ages performed on a series of grammar tests. The focus in the research is the possible effect of gender and age group membership on grammar learning. In collecting the data, we notice that not all students have the same amount of vocabulary knowledge. If we measure this preexisting difference in vocabulary knowledge, we can adjust the grammar test scores by taking vocabulary knowledge into account. This statistical adjustment controls for preexisting differences in a variable which is not the focus of the research and the variable is not deleted from the study, and since it is not the focus of the study, its effect is neutralized by ANCOVA. Therefore, it is possible statistically to control for the effect of a variable by adjusting for preexisting differences in a variable through ANCOVA.

Sample ANCOVA text

A one-way between-groups analysis of covariance was conducted to compare the effectiveness of two different techniques in learning grammar. The independent variable was the type of intervention (deductive, inductive), and the dependent variable consisted of scores on the Grammar Test administered after the intervention was completed.

Participants' scores on the pre-intervention administration of the Grammar Test were used as the covariate in this analysis. Preliminary checks were conducted to ensure that there was no violation of the assumptions of normality, linearity, homogeneity of variances, homogeneity of regression slopes, and reliable measurement of the covariate. After adjusting for pre-intervention scores, there was no significant difference between the two intervention groups on post-intervention scores on the knowledge of grammar, F (1, 30) = .80, p = .12, partial eta squared = .002.

MANOVA

When we have several dependent variables, we use **MANOVA**. For any ANOVA, there is a MANOVA. Therefore we have one-way MANOVA, two-way MANOVA, etc. In one way MANOVA (M + ANOVA), we have one independent variable and two or more dependent variables. The letter M in MANOVA stands for multivariate. In two- way MANOVA, we have two independent variables and two or more dependent variables. Same is true for the analysis of covariance; for any ANCOVA, there is a MANCOVA which can be used for situations where the researcher has data on two or more dependent variables.

A point which needs to be considered is that the dependent variables in MANOVA should be related, or there should be some conceptual reason for analyzing them together. MANOVA informs us about whether there is a significant difference between the groups in a study on a composite dependent variable (a combination of each of your original dependent variables) and also provides the univariate results for each of your dependent variables separately.

Conducting several ANOVAS instead of a MANOVA

A question that sometimes researchers ask is "Can't we just conduct a series of ANOVAs separately for each dependent variable?" The answer is that by conducting a series of analyses, you run the risk of a Type 1 error which means rejecting the null hypothesis by mistake. In fact, the more analyses a researcher runs, the more likely he is to find a significant result, although there might not be any differences between the groups. Therefore, the advantage of using MANOVA is that it decreases the risk of rejecting the null hypothesis by mistake.

Sample MANOVA text

A one-way between-groups multivariate analysis of variance was conducted to investigate education differences in anxiety levels. Three dependent variables were used: state anxiety, trait anxiety and situation anxiety. The independent variable in the study was education. Preliminary assumption testing was conducted to check for normality, linearity, univariate and multivariate outliers, homogeneity of variance covariance matrices, and multicollinearity, with no serious violations noted. There was a statistically significant difference between educated and uneducated on the combined dependent variables, F (3, 360) = 2.43, p = .009; Wilks' Lambda = .83; partial eta squared = .04. When the results for the dependent variables were considered separately, the only difference to reach statistical significance, using a Bonferroni adjusted alpha level of .017, was perceived stress, F (1, 360) = 8.34, p = .004, partial eta squared = .04. An inspection of the mean scores indicated that educated adults reported higher levels of state anxiety (M = 32.42, SD = 3.24) than uneducated (M = 21.02, SD = 2.03).

Unit 15

Statistics: Assumptions of Statistical Tests

Assumptions of Statistical Tests

Most of the statistical tests which were explained in the previous chapter are based on a set of assumptions. When these assumptions are violated, the results of the analysis can be misleading or completely erroneous. The typical assumptions include 1) **normality**, which means the data should have a normal distribution or should be at least symmetric; 2) **homogeneity of variances**, which means the data from several groups have the same variance; 3) **linearity,** which means data should have a linear relationship; 4) **independence**, which means data are independent and 5) **random sampling**, which means the data should be collected randomly. Each will be explained briefly below.

One of the important assumptions for most statistical tests is **normal distribution.** A normal distribution of data means that most of the scores in a set of scores are close to the "average or mean," while relatively few examples are too high or too low. For example, if you look at people's typical daily calorie consumption, you see that most people's consumption will be close to the mean, while fewer people eat a lot more or a lot less than the mean. This is an example of normal distribution. Descriptive statistics, figures, together with certain tests such

Kolmogorov-Smirnov test help you to check whether the data have normal distribution.

Another important assumption that should usually be met for some statistical tests such as ANOVA is the assumption of **homogeneity of variance**s. In order to check this assumption, you should look at the **Levene's test**. Levene's test is used to asses if the groups have equal variances. This test should not be significant to meet the assumption of equality of variances.

Certain tests such as regression require that there should be a linear correlation between the dependent and independent variables. This assumption is called **linearity**. Linearity can be tested graphically using scatter diagrams or via other techniques.

The independence of the data refers to the fact that the data should be collected randomly and independently of data previously selected. For example, repeated measurements from the same people are not independent.

Random sampling is another assumption that needs to be met. Most of parametric techniques assume that the scores are the results of using a random sample from the population.

There are two points which need to be mentioned here. The assumptions of normality and homogeneity of variances are two important assumptions which need to be met for the independent samples t-test, ANOVA and regression. Many researchers are of the opinion that most statistical procedures are almost robust against most violations. For example, one-way ANOVA is said to be a robust technique against the normality assumption which means it can tolerate data that are non-normal. However, several studies have shown that this is often not the case, and that in the case of one-way ANOVA, unequal group sizes can have a negative impact on the technique's robustness.

All in all, the most appropriate statistical test to use not only depends on the design of your research project but also depends on the characteristics of your data which are known as "assumptions". In published articles, most of the time little information is provided on whether the data satisfy the

assumptions of the statistical techniques used which is an unfortunate phenomenon.

Checking normal distribution

There are two main ways to check normality of your data: 1) through **graphs** 2) through **measures tests.**

1) Checking Normal Distribution Through Graphs

In order to check the normal distribution, you can ask the SPSS to give a histogram or box plot. These graphs were explained in Chapter 2. If you look at the following histograms, you can see that only graph 3 has normal distribution.

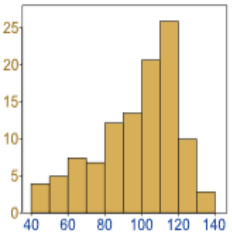


Figure 1

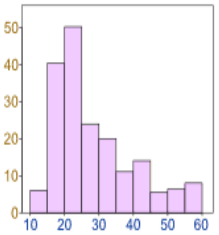


Figure 2

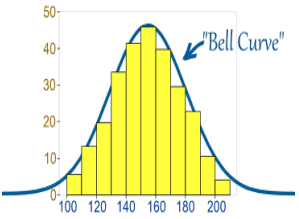


Figure 3

2) Checking normal distribution using measures and tests.

Skewness and Kurtosis

One way to check normality is to look at **skewness** and **kurtosis**. When you ask for descriptive statistics in SPSS, if you wish, it can give you measures of skewness and Kurtosis. These two numerical measures of shape give a more precise evaluation than graphic checking of data for normality. Skewness is a measure of the symmetry in a distribution. A symmetrical dataset will have a skewness equal to 0. So, a normal distribution will have a skewness of 0. Skewness essentially measures the relative size of the two tails. Look at the following figures. Both of them are skewed. That is, none of them is symmetrical. Therefore, when most of the students get high scores, (see Figure 5) or when most of the students get low scores, (see Figure 4), we have the problem of skewedness. If most of the students get high scores, it is negatively skewed and if most of the students get low scores, it is positively skewed. To test the assumption of normal distribution, skewness should be within the range ±2.

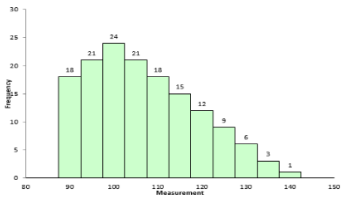


Figure 4: Positively Skewed

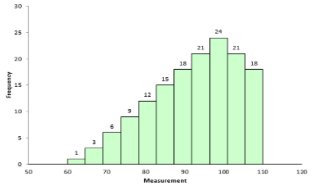


Figure 5: Negatively Skewed

Kurtosis relates to peakedness or flatness of distribution. A flatter distribution has a negative kurtosis. The distribution which is peaked more than a bell shape has a positive kurtosis. Look at Figure 6.

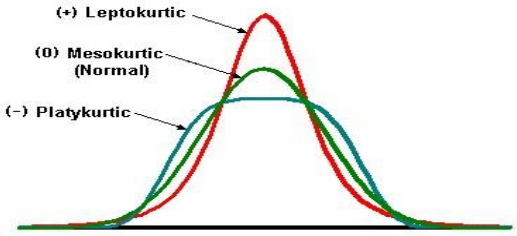


Figure 6: Flatness or Kurtosis of data

As you can see, different types of kurtosis are displayed in the figure. A distribution that is peaked in the same way as any normal distributions is said to be **mesokurtic**. The peak of a mesokurtic distribution is neither high nor low, rather it is considered to be a baseline for the two other classifications.

A **leptokurtic** distribution is one that has kurtosis greater than a mesokurtic distribution. Leptokurtic distributions are identified by peaks that are thin and tall. Leptokurtic distributions are named by the prefix "lepto" which means "skinny".

Platykurtic distributions are those that have a peak lower than a mesokurtic distribution. Platykurtic distributions are characterized by a certain flatness to the peak. The name of this type of distribution comes from the meaning of the prefix "platy" which means "broad". Kurtosis values should be within range of ±7.

Shapiro-Wilk's W Test

Some researchers use **Shapiro-Wilk's W test** to test the assumption of normality. Wilk's test should not be significant to meet the assumption of normality. By significant, I mean when you ask for this test in SPSS it gives you a table entitled "Tests of Normality". You look at the sig value which is the short form of significant value (see right side below Shapiro-Wilk). If it is greater than .05, it means the test is not significant and your data is normal. In the following table, sig

is .074 which is greater than .05. Therefore, the test is not significant and our data have normal distribution.

Table 1

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Pre-test	.138	56	.010	.962	56	.074

Kolmogorov-Smirnov Test

Some researchers use **Kolmogorov-Smirnov test** to test the assumption of normality. This test also should not be significant to meet the assumption of normality.

Checking the Homogeneity of Variances

According to this assumption, all groups should have the same or similar variance. To test the assumption of homogeneity of variance, **Levene's test** is used. Levene's test is used to asses if the groups have equal variances. This test should not be significant to meet the assumption of equality of variances. Levene's test produces an F statistic, the p value of which should be greater than .05. If the p value is less than .05, it

means that that variances of the groups are not equal and it is better to conduct a non-parametric statistical technique.

Parametric vs Non-Parametric Tests

The statistical techniques are classified into two main groups: **parametric** and **non-parametric**. Parametric statistics are more powerful than non-parametric techniques (they are more likely to find a difference or relationship if there are any), but they make assumptions about the data that need to be checked before you conduct one of those techniques. Some of the statistical tests that have been mentioned so far such as independent samples t-test, paired-samples t-test, one-way between-groups ANOVA, and one-way repeated–measures ANOVA need particular conditions to be met, the most important of which are normal distribution, equality of variances and randomization. These types of tests are called parametric tests because normal distribution is one of the characteristics of the population and in statistics we refer to the characteristics of the population as parameters. Therefore, a statistical test which requires normal distribution of data is called a parametric test.

Violation of these assumptions might affect the conclusion of the research and interpretation of the findings. Therefore all studies, whether for a journal article, thesis, or dissertation, must follow these assumptions for accurate interpretation.

In sum, if your data do not meet these requirements, don't worry because there are other statistical tests which you can use to analyze the data. Those tests are called **non-parametric tests**. Non-parametric tests are sometimes called distribution-free tests because they are based on fewer assumptions (e.g., they do not assume that the data is normally distributed). Non-parametric procedures can be used in the following situations:

1) Data do not meet the assumptions of the parametric techniques such as normal distribution.
2) Data are measured on nominal (categorical) scales.
3) Data are measured on ordinal or ranked scale.
4) The sample is very small.
5) There are outliers in the data.

The main problem with using non-parametric tests is that they are generally less **powerful** than their parametric counterparts. It means that they may be less likely to reject the null

hypothesis when the null hypothesis has to be rejected. In the following table, the parametric tests with their non-parametric equivalents are provided. A final note to make here is that the two types of chi-square tests mentioned before are considered to be among the non-parametric tests. The Table 2 displays the parametric tests with their non-parametric equivalents.

Table 2

Parametric and Non-parametric Tests

Parametric Test	**Non-parametric Test**
Independent samples t-test	Mann-Whitney U Test
Paired-samples t-test	Wilcoxon Signed Rank Test
One-way between-groups ANOVA	Kruskal-Wallis Test
One-way repeated –measures ANOVA	Friedman Test

Unit 16

Statistics: Effect Size

Effect Size

The elements of hypothesis testing include 1) stating the null hypothesis, 2) stating the alternative hypothesis, 3) selecting a level of significance 4) collecting and analyzing the sample data, 5) referring to a criterion for evaluating the sample evidence, 6) rejecting or failing to reject the hypothesis and finally, 7) calculating **the effect size**. Unfortunately some researchers skip the effect size step.

Many researchers conduct statistical tests and when they see that the significant value is less than .05, they conclude that there is a difference between the groups or there is a relationship between two variables. They ignore the fact that although the findings are **significant**, they are not **important**. In other words, there might be a relationship or difference there, but the amount or size of that relationship or difference is not large enough to be considered important.

A result that is **statistically significant** may lack **practical significance** because there is a direct relationship between the size of the sample and probability of rejecting a false null hypothesis. Sometimes a researcher fails to reject the null hypothesis because the sample is not large enough, but the

effect size is great and sometimes since the sample is large, a very small difference between groups which means a very small effect size might become significant. Therefore, “significant” and “important” or “statistically significant” and “practically significant” are two different issues. Keep in mind that your findings might be significant without being important. Therefore, whenever you conduct a test such as a t-test, an ANOVA, a correlation test, a chi-square test, etc., you had better look at the effect size of the results too.

Effect size refers to the magnitude of the impact of the independent variable on the dependent variable. Effect size gives researchers information about whether the difference between groups is important or not. In fact, an effect size enables the researcher to see the size of the difference between the groups. If the effect size is small, then it is reasonable to consider the findings as unimportant, even if they are statistically significant. If the effect size is large, then the researcher has found something that is important to report to other researchers.

Different Types of Effect Sizes

In general, we can divide the effect sizes into two categories: **group difference effect sizes** and **relationship effect sizes**. Group difference effect sizes refer to the difference between the means of two groups. For example, after calculating the t-test on two means, we also determine the size of the difference between two means. The most common effect size in this category is **Cohen's d**. Cohen's d shows the difference between two means in standard deviations. For example, if Cohen's d is 2, it means that groups differ by two standard deviations.

Relationship effect sizes show how much the independent and dependent variables vary together. In this type of effect size, the more two variables are related, the higher the effect size is. This type of effect size indicates the proportion of variance of the dependent variable that is explained by the independent variable. The most common types of effect sizes in this category are **eta- squared**, **omega squared**, **partial eta-squared** and **partial omega squared**. Therefore, based on the above-mentioned points, effect size refers to the strength of the difference between groups, or the strength of the influence of the independent variable on the dependent variable.

A question that you may ask at present is which one is easier to understand and which one is better to use. To answer these questions, you should remember that group difference effect size or Cohen's d can go higher than 1. In other words, it ranges from 0 to any number, but correlation effect sizes range from 0 to 1 and as a result it is easier to compare effect sizes which are all between 0 and 1. Some effect sizes are more associated with certain statistical tests. For example, for correlation and regression, correlation effect sizes which include eta-squared, and omega squared are more commonly used while Cohen's d is more common with ANOVA and t-test.

Interpreting an Effect Size

After you calculate the effect size, you should decide whether it is small, medium or large. The following tables, which are based on Cohen (1988), show you the criteria to make such decisions. Table 1 shows the criteria based on Cohen's d. Table 2 shows the criteria based on eta-squared and finally, Table 3 shows the effect size with regard to correlation coefficient.

The following table shows group difference effect sizes in standard deviation terms.

Table 1

Cohen's d Effect Sizes

Size	Cohen's d
Small	.2
Medium	.5
Large	.8

For relationship effect sizes, again the effect size can be small, medium or large. The following table (Table 2) shows the criteria.

Table 2

Eta-squared Effect Sizes

Size	Eta-squared
Small	.01
Medium	.06
Large	.14

As to the correlation, there are two ways to decide whether the effect size is small, medium or large. One way is to simply look at the size of *r* and compare it with Table 3.

Table 3

Values of r Effect Sizes

Size	Eta-squared
Small	.10 to .29
Medium	.30 to .49
Large	.50 to 1

Another way is to calculate the **coefficient of determination** to see how much variance your two variables share. All you need to do is to square your r value (multiply it by itself) and then multiply it by 100. For example, two variables that correlate r=.3 share .3 × .3 = .09 = 9 percent of their variance. There is not much overlap between the two variables. A correlation of r=.6, however, means 36 percent shared variance (.6 ×.6 = .36).

References

American Psychological Association. (2010). *Publication manual of the American Psychological Association (7th ed.)*. Washington, DC: Author.

Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches*. Los Angeles: Sage.

Dörnyei, Z. (2007). *Research methods in applied linguistics.* Oxford: Oxford University Press.

Perry, F. (2005). *Research in applied linguistics: Becoming a discerning consumer*. Lawrence: Erlbaum Assoc Inc.

Mackey, A., & Gass, S. M. (2005). *Second language research: Methodology and design.* Mahwah, NJ: Erlbaum.

McKay, S.L. (2006). *Researching second language classrooms*. Mahwah, NJ: Lawrence Erlbaum.

By Mohammad Golshan

Website: www.nokhbeganfarda.com

1. A Dictionary of Persian-English Idioms
2. Practical English Sentences in Everyday Life (Volume 1)
3. Practical English Sentences in Everyday Life (Volume 2)
4. Key Idioms in Key Sentences
5. The Most Frequent Questions Students of English Ask
6. Listening Comprehension of English Movies
7. Understanding the Expressions Used in American Movies
8. Speak English Like an American (Translated book)
9. A Dictionary of the Origins and Stories of English Idioms
10. Idioms and Their Stories (Translated book)
11. TOEFL Idioms (Translated book)
12. Street Talk 1/2/3 (Translated book)
13. Speak English Easily on a Trip
14. Practical Words, Phrases and Sentences in English Law Texts
15. Key Sentences in English Law Texts
16. The English Idioms You Need to Know
17. Success in IELTS Writing
18. Success in IELTS Speaking
19. Key English Sentences for Conversation
20. Essential Grammar for the TOEFL
21. Essential Words for TOEFL-IELTS-MCHE-TOLIMO-EPT
22. Common Collocations in American English
23. English Question Bank for Conversation
24. Listening Skills Through Films
25. Common Proverbs in English

26. Idioms Through Pictures
27. IELTS and TOEFL Vocabulary Through Pictures
28. Funny Riddles and Quotations in English
29. 504 Essential Grammar Points You Need to Know
30. 524 Interesting and Strange Points About English
31. Teach English to Your Children at Home
32. Differences Between English Words in Simple Language
33. New Words and Phrases in English
34. Listening Comprehension of Real American English
35. Key Sentences for Letter Writing
36. Common Words and Phrases in American Movies
37. Practical Prefixes, Suffixes and Roots in English
38. What you Need to Know About Prepositions in English
39. 421 Useful Words and Phrases in English Films
40. Speak Business English Like an American (Translated book)
41. Speak English Around Town (Translated book)
42. More Speak English Like an American (Translated book)
43. All you Need to Know About Articles a/an/ the
44. English Phrasal Verbs You Need to Know
45. Common Phrasal Verbs in English
46. Common Idioms in American English
47. An English Course for the Students of Management
48. Golshan Comprehensive Dictionary of English-Persian Idioms
49. What You Need to Write an Article in English
50. A Guide to Writing and Presenting Articles in English
51. Successful Presentation of an Article in English
52. A Guide to Editing Scientific Papers

53. Article Writing Made Easy
54. Proposal Writing Made Easy
55. Essential Words and Phrases for Writing Articles in English
56. Statistics of Theses and Articles in Simple Language
57. Language Testing Made Easy
58. Essential Texts for Writing Theses in English
59. Essential Templates for Writing Articles in English
60. Common Academic Phrases in English
61. Corrective Feedback: From Theory to Practice
62. An English course for the Students of Nursing
63. An English course for the Students of Psychology
64. An English course for the Students of Midwifery
65. Words of Language Exams in One Book
66. Academic Phrases Every Researcher Needs to Know
67. Templates for Writing Proposals
68. Key Collocations Your Need to Know
69. Free Discussion in English Classes
70. Common Words and Phrases in Legal Deeds and Correspondence
71. Principles of Language Teaching
72. English for Language Classes
74. Creating Motivation in Language Classes
75. Teaching English to Young Learners
76. Proverbs, Idioms and Phrasal Verbs for English Teachers