Statistical Analyses and Interpretation of Data

Mulugeta Gebreselassie*

Introduction

Any statistical investigation involves four important stages: data collection, data organization, data analysis and interpretation or inference. This deals with one aspect of the four main stages of statistical investigation.

Before data analysis and interpretation it is mandatory to devise procedures of collecting "quality" data. If there is any problem with the quality of the data it is unthinkable to obtain precise results by applying sophisticated methods of analysis.

A pre-requisite to obtain quality data is a good design of the data collection instrument. This includes a good questionnaire design and a good sampling design. Though the author feels that the prerequisite topics are covered in previous training sessions, a mention of a sampling design is relevant.

The main theme of this topic, analysis and interpretation of data is based on the concept of a sample (a portion of a population selected systematically for a pre-determined purpose). A sample is a compromise between cost and precision. To know the precision of estimates, the system of selection should be random (random selection is a selection technique, which gives every unit in the population a known non-zero chance of being included in the sample). In fact, the stages of designing a questionnaire and sampling should be administered in consultation with professionals, because poor work in one phase may ruin a study or survey in which everything else is done well and calling the statistician at the end, unless he is particularly gifted or lucky, will only make him able to provide a post mortem report on the reasons why the study failed to attain its goals.

* Lecturer, Department of Statistics, Addis Ababa University.

Statistical Analyses and Interpretation of Data

Mulugeta Gebreselassie*

Introduction

Any statistical investigation involves four important stages: data collection, data organization, data analysis and interpretation or inference. This deals with one aspect of the four main stages of statistical investigation.

Before data analysis and interpretation it is mandatory to devise procedures of collecting "quality" data. If there is any problem with the quality of the data it is unthinkable to obtain precise results by applying sophisticated methods of analysis.

A pre-requisite to obtain quality data is a good design of the data collection instrument. This includes a good questionnaire design and a good sampling design. Though the author feels that the prerequisite topics are covered in previous training sessions, a mention of a sampling design is relevant.

The main theme of this topic, analysis and interpretation of data is based on the concept of a sample (a portion of a population selected systematically for a pre-determined purpose). A sample is a compromise between cost and precision. To know the precision of estimates, the system of selection should be random (random selection is a selection technique, which gives every unit in the population a known non-zero chance of being included in the sample). In fact, the stages of designing a questionnaire and sampling should be administered in consultation with professionals, because poor work in one phase may ruin a study or survey in which everything else is done well and calling the statistician at the end, unless he is particularly gifted or lucky, will only make him able to provide a post mortem report on the reasons why the study failed to attain its goals.

* Lecturer, Department of Statistics, Addis Ababa University.

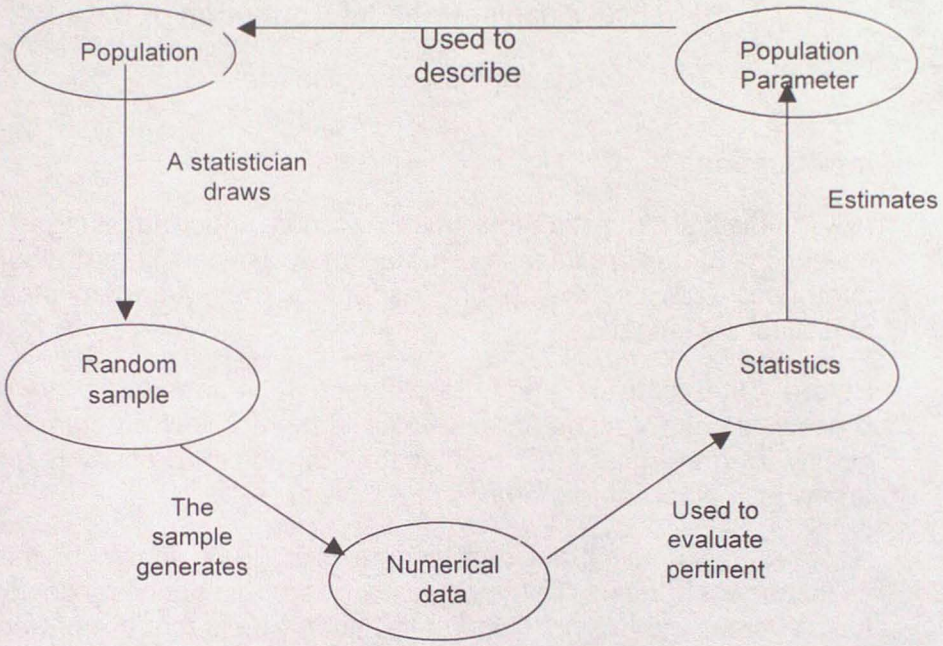

Fig. 1: Summary of a Statistical Process

Analysis of Data

Once data have been collected they have to be analyzed. Three factors which affect our choice of methods of data analyses are:

- the number of variables being examined
- the level of measurement of the variables
- whether we want to use our data for descriptive or inferential purposes.

The number of variables

How we analyze data depends on what we want to find out. If we simply wish to describe the characteristics of the sample at a time, we use a univariate method of analysis. If we are interested in two

variables simultaneously, we use bivariate methods. Similarly, if our research question makes use of three or more variables, we use multivariate techniques.

Example 1: The percentage of girls in AAU is 2%.
Example 2: If we want to see if gender differences bring about differences in educational achievement of pupils, we use bivariate methods.
Example 3: We might be interested in the determinants of school enrolment like age, sex, household income, etc.

Before analyzing data we must be clear about the question we are trying to answer. This dictates the broad type of analysis we choose. Once we have decided, for example, that we need to use a bivariate technique, we then need to choose between a range of such techniques. In practice we develop and refine our research question in the process of analysis also we move between univariate, bivariate and multivariate techniques.

Example: Initially we might formulate a question "do pupil's enrolment status vary because of their age?" Bivariate analysis might show that they do tend to an extent but that we need to look at some other variables together with age. That leads to using multivariate techniques.

Level of Measurement

In the choice among univariate, bivariate, and multivariate techniques, the level of measurement is a key factor. Level of measurement refers to the categories - nominal, ordinal, interval and ratio data. The presence or absence of these levels can be somehow affected by decisions of the researcher. So which level is the best to aim for?

- a wider range of analysis is appropriate as the level of measurement increases.
- more powerful and sophisticated techniques of analysis are appropriate for interval variables.

Generally it is advisable to measure variables at the highest level appropriate to that variable, but considerations of reliability, response rate and need will mean that measurement at lower levels often makes most sense.

Table 1: Examples of Univariate, Bivariate and Multivariate Methods

Univariate methods	Bivariate methods	Multivariate methods
• Frequency distribution	• Cross-tabulation	• Customized tables
• Descriptive measures	• Scatter graphs	• Partial correlation
	• Regression	• Multiple correlation
	• Correlation	• Multiple regression
	• Comparison of means	• MANOVA
		• Discriminant analysis

Using these methods the researcher organizes his/her data, or extracts relevant information. Information extraction involves:

- studying how individual pieces of information cluster together (measures of central tendency);
- studying how individual cases spread apart (measures of dispersion).

Consider a survey of 500 persons or an analysis of 500 questionnaires. If only ten questions are asked of each person or each questionnaire, such a study produces at least 500 pieces of information. The hundreds of answers to each question, in their raw form, are very difficult, if not impossible to interpret without the help of some summary measures or statistics. This becomes more and more the case as the size and the complexity of a set of data increases.

Statistical Inference (Interpretation of data)

Statistical inference is a statistical process of drawing valid conclusions about the population under investigation from a described sample data. These can be done in two different procedures:

parametric and nonparametric methods. In the former case scrutiny of several statistical assumptions such as normality, independence and homogeneity of variances are mandatory. But in the later case there is no need to demonstrate the validity of these assumptions, except the assumption of continuity.

There are two broad categories of the problems of statistical inference, estimation and hypothesis testing. In the case of estimation the investigator will have no pre-conceived notion about the value of the parameter under question. He/she will be going to the investigation with the intent of answering the question, what is the numerical value of the parameter? But in a hypothesis-testing problem the investigator usually has a prior notion as to the value of the parameter. The purpose of the study is to gather evidence that either affirms or refutes the hypothesized theory.

Estimation

- *Point estimation:* It is the process of identifying a single figure that gives information about a parameter.
  Example: The average age of the trainees is 30.

- *Interval estimation:* It is the process of identifying the upper and lower limits of an interval that gives information about a parameter with some degree of confidence probability.

  Example: I am 95% confident that the average age of the trainees lies in the interval (25, 35).

Further examples

- We are always 95% sure that the population mean lies between ($\bar{x}$-1.96σ/n and $\bar{x}$+1.96σ/n)
- We are always 99% sure that the population mean lies between ($\bar{x}$-2.58σ/n and $\bar{x}$+2.58σ/n)

Hypothesis testing

A researcher has always some fixed ideas about certain parameters based on prior experiments, surveys or experience. However, these have to be ascertained for their validity by collecting information in the form of sample data. In this case we usually try to answer the following questions:

- How strong is the relationship (difference) between variables?
- Is this relationship (difference) real or is it due to chance?

To answer these questions a statistical procedure called " Statistical test", which is governed by probability rules to take a decision about the hypothesized questions, is set. In the course of doing these, two types of errors occur, type I and type II errors.

Type I error – an error committed by rejecting a true hypothesis
Type II error – an error committed by accepting a false hypothesis

Generally, the following steps are important in any hypothesis testing exercise:

1st formulation of the hypothesis
2nd fixing the level of significance
3rd choosing an appropriate test statistic
4th making statistical decision (rejecting or accepting the hypothesis)
5th making inference (drawing valid conclusion)

Inference about the mean

Let $x_1, x_2 \ldots x_n$ be a sample of size n from a population which is normally distributed with mean μ and variance δ^2. The possible types of hypotheses that can be raised are:

- Univariate case

  H_0: $\mu = \mu_0$ H_0: $\mu = \mu_0$ H_0: $\mu = \mu_0$

  H_1: $\mu \neq \mu_0$ H_1: $\mu > \mu_0$ H_1: $\mu < \mu_0$

- Bivariate case

  H_0: $\mu_1 = \mu_2$ H_0: $\mu_1 = \mu_2$ H_0: $\mu_1 = \mu_2$

  H_0: $\mu_1 \neq \mu_2$ H_0: $\mu_1 > \mu_2$ H_0: $\mu_1 < \mu_2$

- Multivariate case

  H_0: $\mu_1 = \mu_2 = \ldots = \mu_n$

  H_1: at least one pair is different

And similar hypotheses can be listed down for a population proportion, which is an important parameter in social science research. For the mean as well as the proportion, according to the type of hypothesis and structure of the population different test procedures are suggested. The following are some of the procedures.

The Z-test

It is used in the case of a normally distributed population with known variance or when the sample size is very large.

Example: Suppose an educator decided to check if the average performance of his students was 72.4. He took a sample of 35 students. He found out that the average score was 73.2. Do you think he would reject or accept his hypothesis? (δ=2.1 and α = 5%)

- H_0: $\mu = 72.4$

  H_1: $\mu \neq 72.4$
- α= 5%

Since n > 30 we can use the Z - test, $Z_{cal} = 2.25$ and $Z_{0.025} = 1.96$

- Since $Z_{cal} > Z_{0.025}$, H_0 is rejected.
- At 5% level of significance the average performance of the students in the class is different from 72.4

The t-test

In case the sample size is small (n<30) and the population variance is unknown it is the best test statistic.

	Urban	Rural
Mean	65	58
Variance	170	170

Example: An educator suspected that rural students' mathematics performance was better than urban students. He took a sample of 18 urban and 14 rural students and found the following. Do you think his suspicion would be true at α=5%?

- H_0: $\mu_U < \mu_R$
- H_1: $\mu_U < \mu_R$
- α= 5%
- Since $n_1 < 30$ & $n_2 < 30$ use the t-test, $t_{cal} = 1.506$, $t_{0.05}(30) = 1.96$
- H_0 is not rejected
- Urban and rural students have similar performance in maths at 5% level of significance.

Note: in making inference in the bivariate case

- if $\delta^2_1 \neq \delta^2_2$, and δ^2_1& δ^2_2 are known, use the Z-test
- if $\delta^2_1 \neq \delta^2_2$, but δ^2_1& δ^2_2 are unknown, use the variants of the t-test

Example: To compare literature and science students with respect to their interest in modern music an aptitude test was given to 60 literature and 75 science students and the following was found.

	Literature	Science
Mean	76.4	81.2
Standard deviation	8.8	6.2

Do the data give evidence on the difference in interest of both groups at 95% confidence level?

- H_0: $\mu_U < \mu_R$
- H_1: $\mu_1 \neq \mu_2$
- α= 5%
- The appropriate test statistic is t, t_{cal} = 3.57 and $t_{0.025}$ (133) = 2.326
- H_0 is rejected.
- At 5% level of significance the data gives sufficient evidence to conclude that literature and science students have a difference in their music interests .

ANOVA

ANOVA is an abbreviation for Analysis of Variance. It is used to make inference about the equality of three or more means. It is a technique of making inference by decomposing the total variance into *between groups variance* and *within group variance*. It uses the F-test as a test statistic.

- H_0: $\mu_1 = \mu_2 = \mu_3 \ldots = \mu_k$
- H_1: at least one of the pairs is different
- α= 5%
- The appropriate test statistic is F as shown in the ANOVA table

ANOVA Table

Source of variation	Df	Sum of squares	Msq	F
Between groups	k-1	SSB	MSB	MSB/MSW
Within groups	n(k-1)	SSW	MSW	
Total	nk-1	SST		

Chi-square

Relationships between two or more variables can be studied by using correlation, regression and Chi-square test. When the variables of interest are continuous and normally distributed, correlation and regression are appropriate, but if the variables are categorical in

nature Chi-square test is more appropriate. The Statistic used in these tests is:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Where O_i = Observed frequency in the category
E_i = Expected frequency in the category

The distribution of χ^2 is chi-square with degrees of freedom (k-1) is used as long as the sample is not too small. Chi-square measures the concordance between the observed and the expected frequencies. Thus, a small value of χ^2 implies agreement while large values of χ^2 implies disagreement between the observed and expected frequencies.

The Chi-square distribution can be used to test

- Goodness of fit
- Independence of attributes
- Homogeneity of populations

Test of Independence

Here the Chi-square test procedure is used to test the hypothesis of independence of two attributes, i,e.

H_o: the two attributes are independent
H_1 : the two attributes are associated

Education level	Gender				Total
	Male		Female		
	Observed	Expected	Observed	Expected	
Primary	O_{11}	E_{11}	O_{12}	E_{12}	$n_{1.}$
Secondary	O_{21}	E_{21}	O_{22}	E_{22}	$n_{2.}$
Tertiary	O_{31}	E_{31}	O_{32}	E_{32}	$n_{3.}$
Total	$n_{.1}$		$n_{.2}$		n

Suppose two variables (categorical) have ***r*** and ***c*** mutually exclusive classes or attributes respectively. For instance if we consider gender

(male, female) and level of education (primary, secondary, tertiary): gender has two classes while level of education has three classes.

The set of data that can be observed from a sample of size ***n***.. can be presented as indicated in the table above.

In the computation of the Chi-square statistic the expected frequencies (E_{ij}) are computed as:

$$E_{ij} = \frac{n._j \, x n_i.}{n..}$$

And, therefore Chi-square is given by:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2 \text{ ((r-1)(c-1))}$$

Example: A sociologist studying family size wondered if size of a family was related to the level of education attained by parents. A survey on 254 couples in the age group from 35 to 50 who had reportedly completed their education was taken. The number of children was recorded along with the highest level of education attained by either parent as shown.

No.of children	Level of Education				Total
	Postgraduate	College graduate	High School Graduate	Didn't complete high school	
6*	6 (15.2)	13 (17.9)	14 (8.3)	17 (8.7)	50
3-5	12 (13.3)	16 (15.8)	8 (7.3)	8 (7.6)	44
1-2	38 (28,8)	40 (34.0)	11 (15.7)	6 (16.5)	95
0	21 (19.7)	22 (23.3)	9 (10.7)	13 (11.2)	65
Total	77	91	42	44	254

If we are asked to test whether both variables are related or not, we can use Chi-square as follows:

H_o: Family size and level of educational attainment are independent

H_1: Not H_o

The calculated χ^2 is 31.76 while the table value is $\chi^2_{0.05}$= 16.92

Conclusion: Family size is associated with level of educational attainment of parents.

Test of Homogeneity

Suppose we want to see if populations $A_1, A_2, \ldots A_r$ are homogenous with respect to a certain characteristic say B with C categories. If we select $n_{i.}$ units from each population and if each sample of size $n_{i.}$ is classified into C categories say $B_1, B_2, \ldots, B_c$ then we get the following type of tabular representation.

Population	Characteristic B			
	B_1	B_2	B_c	Total
A_1	O_{11}	O_{12}	O_{1c}	$n_{1.}$
A_2	O_{21}	O_{22}	O_{2c}	$n_{2.}$
A_r	O_{r1}	O_{r2}	O_{rc}	$n_{r.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.c}$	$n_{..}$

The hypothesis of interest in this case will be

H_o: $P_{1j}=P_{2j}=\ldots=P_{rj}$ for all j

H_1: not H_o

The test statistic that can be used here is

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2 ((r-1)(c-1))$$

Example: To make a study of market penetration of a cosmetic item the marketing division of a company selected a random sample of 270,250 and 300 consumers from three occupation categories and obtained the following.

Occupation	Market penetration			Total
	Never heard about the cosmetic	Heard about It but not buy it	Bought it at least once	
High school	45 (53.9)	62 (79.7)	163 (131.4)	270
College girls	80 (54.6)	95 (73.8)	75 (121.6)	250
Employed Women	54 (65.1)	85 (88.5)	161 (146.0)	300
Total	179	242	399	820

Do these data indicate that the extent of market penetration differs in the three occupation categories?

The computed $\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ = 52.46 and the table value at 4 degrees of freedom is 9.49, hence the hypothesis is rejected.

Conclusion: The product did not penetrate the markets of the three occupation groups equally.

Correlation and Regression

The methods of analysis discussed so far allow us to workout whether two or more variables are associated: whether people who vary on one variable also vary systematically on the other. How strongly are these variables associated?

Correlation and regression are statistical procedures, which help to identify the strength, direction and nature of relationship between two or more variables.

Example: We may want to know how much someone with a given amount of education is likely to earn, particularly how much difference will staying on at school for another year make someone's income level?

Correlation will tell us how likely more education is to affect income. But regression analysis tells us how much difference it is likely to make. It enables us to make predictions about the dependent variable for fixed values of the independent variables. It also enables us to say how much impact each unit change in the independent variable has on the dependent variable. Depending on the number of explanatory variables available we can have simple or multiple correlation and regression.

Note

Originally presented to CERTWID in December 2000 for training of researchers.