Unit 15

Sampling Methods and Estimation of Sample Size

Contents

Learning Objectives

It is expected that after reading Unit 12 you would be able to

- ❖ Define what is sampling
- ❖ Classify sampling methods
- ❖ Calculate sample size.

15.1 Introduction

Unit 15 deals with the procedure of sampling© that helps you arrive at a subset of the universe of your research. It discusses the various methods of sampling and tells you how to work out a sample size. You will again read about sampling in Block 6. This is a subject you will need to master carefully as no matter what type of research you wish to carry out, you will need to apply your skill of the craft of sampling.

15.2 Sampling

A sample is a subset of the population that represents the entire group. When the population (or universe) is too large for the researcher to survey all its members because of its cost, the number of personnel to be employed, or the time constraint, a small carefully chosen sample is extracted to represent the whole (see Figure 15.1). The sample, as drawn in Figure 15.1, is expected to reflect the characteristics of the population.

A well selected sample may provide superior results. For example, in a research where well-trained interviewers are required, it may be possible to get a few trained interviewers to collect a sample rather than to get many trained interviewers to investigate the entire population. The trained interviewers may gather better quality information than non-trained or less trained interviewers. By contrast, if the population is sufficiently small, the entire population should be studied. When data are gathered on each and every member of the population, the study is known as a census study. The researcher is expected to clearly define the target population.

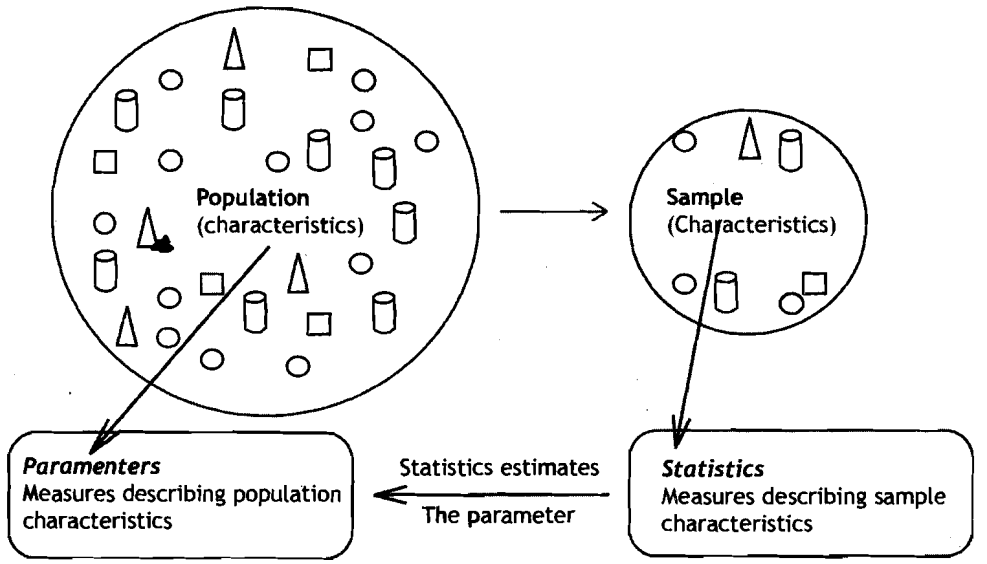


Figure 15.1

Relationship between Population, Parameter, Sample and Statistics

A population may be defined as an aggregate of individuals possessing a common trait or traits. There are two important factors: first that a population is the complete group about which knowledge is sought, and second each and every individual has some certain specified attribute or attributes.

Let us now complete Reflection and Action 15.1.

Reflection and Action 15.1

Work out the relationship between population, parameter, sample and statistics to reflect the characteristics of the population of the unit of your research project.

15.3 Classification of Sampling Methods

Sampling methods are classified into Probability or Non-probability. If the purpose of research is to draw conclusions or make predictions affecting the population as a whole (as most research usually is), then one must use probability sampling. But, if one is only interested in exploring how a small group, perhaps even a representative group, is doing for purposes of illustration or explanation, then one may use non-probability sampling.

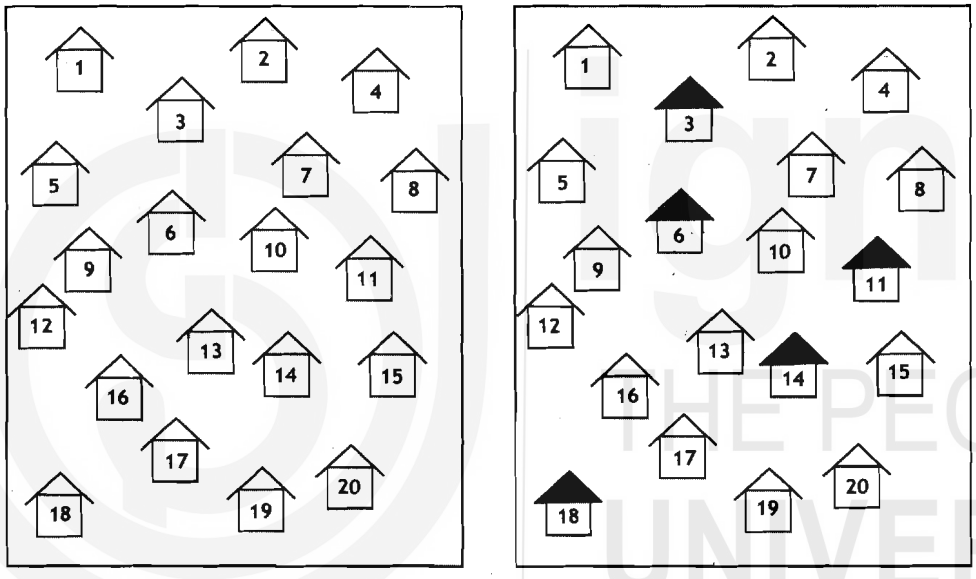
Let us first discuss probability sampling.

(A) Probability Sampling

In probability samples, each member of the population has a known non-zero probability of being selected. The key point behind all probabilistic sampling approaches is random selection. The advantage of probability sampling is that sampling error can be calculated, which is the degree to which a sample might differ from the population. Probability methods include random sampling, systematic sampling, and stratified sampling.

We shall discuss each of them.

a) **Random sampling** is the purest form of probability sampling. Each member of the population has an equal and known chance of being selected. The prerequisite for a random sample is that each and every item of the universe has to be identified. Random selection is effective in a clearly defined population that is relatively small and self-contained. When the population is large, it is often difficult or impossible to identify its each and every member, so the assemblage of available subjects becomes biased. One obtains a list of all residents or the voters list or telephone directory, and then selects a sample using a sequence of numbers from a random numbers table. Random numbers can also be created in numerous computer softwares. See Figure 15.2 that illustrates the selection of sample using random number table.



Population

Simple Random Sample drawn
(dark colour)
using Random number table

Figure 15.2 Population Simple Random Sample drawn (dark colour) using Random Number Table

Source: Fisher, R. A. and F. Yates 1982. Statistical Tables. Longman: New York

b) **Systematic sampling** is also called an "N^{th}-name selection" technique. After the required sample size has been calculated, every N^{th} record is selected from a list of population members. As long as the list does not contain any hidden order, this sampling method is as good as the random sampling method. Its only advantage over the random sampling technique is simplicity. Systematic sampling is frequently used to select a specified number of records from a computer file. In Figure 15.3 you can find elucidation of the systematic random sampling method. The first number (2) has been selected by random number, followed by the selection of every 5^{th} item in the series.

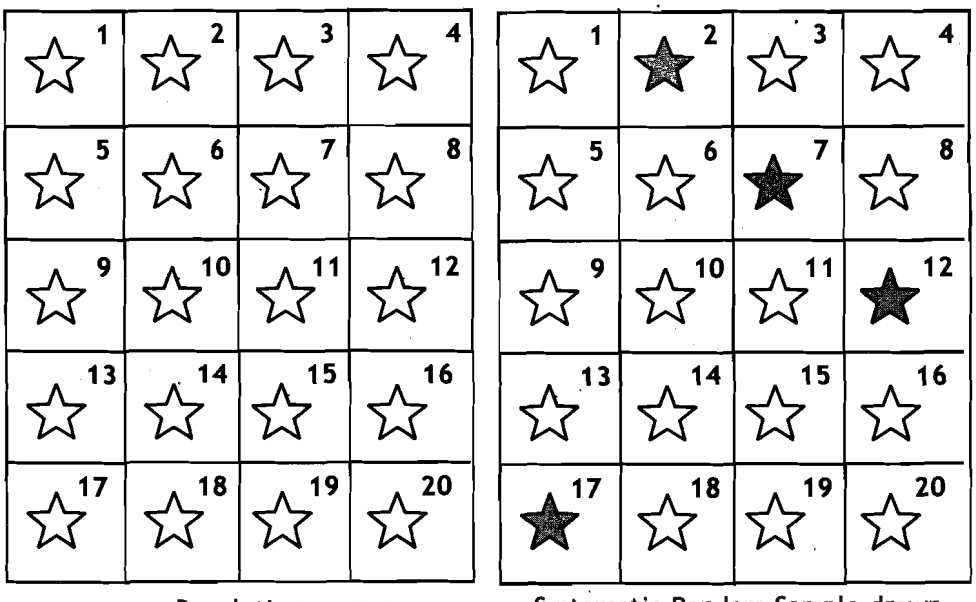


Figure 15.3 Systematic Random Sampling Method

c) **Stratified sampling** is a commonly used probability method that is superior to random sampling because it reduces the sampling error. A stratum is a subset of the population that shares at least one common characteristic. Examples of strata might be males and females, or managers and non-managers. The researcher first identifies the relevant strata and their actual representation in the population. Random sampling is then used to select a 'sufficient' number of subjects from each stratum. 'Sufficient' refers to a sample size large enough for the researcher to be reasonably confident that the stratum represents the population. Stratified sampling is most successful when (i) the within variance of each stratum is less than the overall variance of the population; (ii) when the strata in the population are of unequal size or have unequal incidence; and (iii) when sampling is cheaper in the strata. Figure 15.4 shows stratified random sampling method. Samples from the three strata have been extracted in proportion to their numbers.

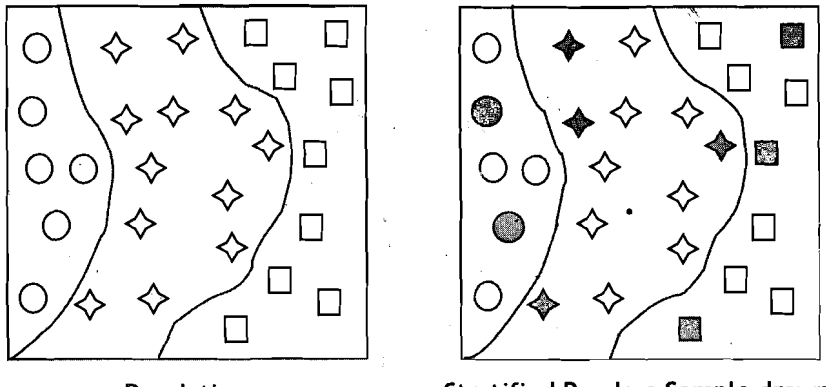


Figure 15.4 Stratified Random Sampling Method

d) **Cluster random sampling** is useful when the population is dispersed across a wide geographic region. This method allows one to divide the population into clusters and then select the clusters at random. Thereafter one can either study all the members of the selected clusters or again take random (simple or systematic) samples of these sampled clusters. If the latter system is followed, it is called multi-stage sampling. This method, for example, could be effective to study a tribal group or a community that is dispersed. The villages could be used as clusters and can be randomly selected. Figure 15.5 shows that five blocks (2, 7, 10 and 14) out of sixteen have been selected by random number. Each block contains a series of samples, as illustrated.

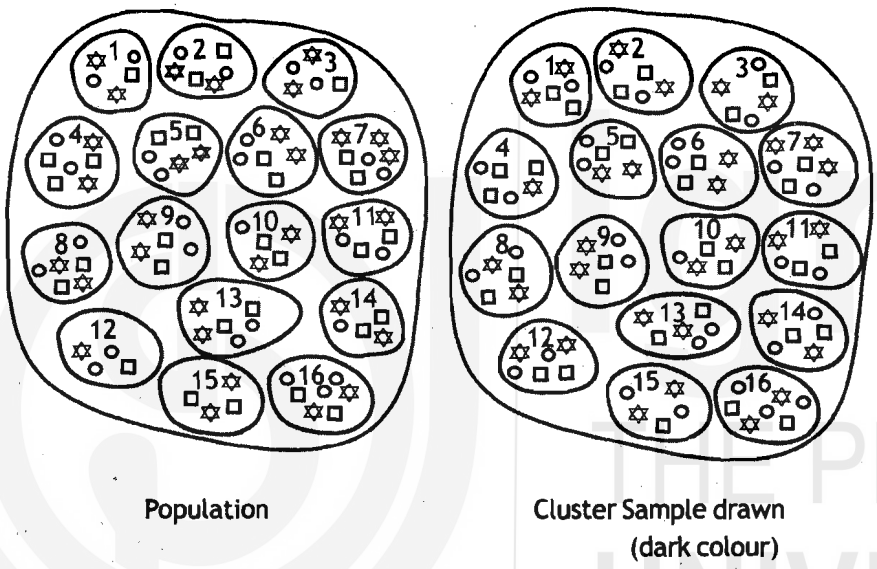


Figure 15.5 Cluster Random Sampling Method

Reflection and Action 15.2

Following the figures in the text, make figures based on the population pertaining to your research project that you selected while computing R & A 13.1 & 13.2 to show

i) population simple random sample drawn in dark colour, using random number table

ii) systematic random sampling method

iii) stratified random sampling method

iv) cluster random sampling method

(B) Non-probability Sampling

In non-probability sampling, members are selected from the population in some non-random manner. In this method, the degree to which the sample differs from the population remains unknown. Non-probability methods include Convenience sampling, Judgment sampling, Quota sampling and Snowball sampling. Let us now discuss each of the non-probability sampling methods.

a) **Convenience sampling** is used in exploratory research where the investigator is interested in getting an inexpensive approximation of the fact. As the name implies, the sample is selected because it is convenient. Also called haphazard or accidental, this method is based on using people who are a captive audience, just happen to be walking by, or show a special interest in research. The use of volunteers is an example of convenience sampling. This method is often used during preliminary research efforts to get a gross estimate of the results, without incurring the cost or time required to select a random sample.

b) **Judgment sampling** is a common non-probability method. The researcher selects the sample based on judgment. This is usually an extension of convenience sampling. For example, a researcher may decide to draw the entire sample from one 'representative' village, even though the population may be distributed over a number of villages. When using this method, the researcher 'feels' that the chosen sample is representative of the entire population.

c) **Purposive sampling**, much similar to judgment sampling, is where the researcher targets a group of people believed to be typical or average, or a group specially picked for some unique purpose. The researcher never knows if the sample is representative of the population, and this method is largely limited to exploratory research.

d) **Quota sampling** is the non-probability equivalent of stratified sampling. Like stratified sampling, the researcher first identifies the strata and their proportions in the population. Then convenience or judgment sampling is used to select the required number of subjects from each stratum. The researcher resorts to haphazard or accidental sampling, and makes no effort to contact people who are difficult to reach. This differs from stratified sampling, where the strata are filled by random sampling.

e) **Snowball sampling** is a special non-probability method used when the desired sample characteristic is rare. It may be extremely difficult or cost prohibitive to locate respondents in these situations. Snowball sampling relies on referrals from initial subjects to generate additional subjects. In other words, snowball sampling comprises identification of respondents who in turn refer researches to other respondents. This technique provides a means to access relatively invisible and vulnerable social groups. While this technique can dramatically lower the search costs, it comes at the expense of introducing bias because the technique itself reduces the likelihood that the sample will represent a good cross-section of the population. For example, an investigator finds a rare genetic trait in a person, and starts tracing his pedigree to understand the origin, inheritance and etiology of the disease.

You may have heard that only quantitative researches require sampling. The fact is that qualitative researches use sampling procedures (see Box 15.1).

Box 15.1 Use of Sampling in Qualitative Research

As Berger (1989) and Sarantkos have pointed out , it is fairly common for qualitative researches to use sampling procedures in the following manner.

i) Sampling is relatively small, dealing with typical cases.
ii) Use of flexible samples in size not requiring statistical calculations
iii) Use of purposive sampling dealing with non-probability
iv) Use of sampling to achieve suitability rather than representativeness
v) Sampling occurs while the research is in progress, rather than selecting a sample before starting it.

We would now focus on the procedure of calculating the sample size.

15.4 Sample Size

A prudent choice of the sample size for a particular survey involves many considerations, among which are the resources in manpower, cost per sample units and funds available, the number and type of parameters to be estimated. Obviously, these specifics will vary from one survey to another. All the same, a framework can be constructed within which general and viable decisions with respect to sample size can be taken. Sampling theory aids in arriving at good estimates of the sample size. The standard error here too provides the key.

Apart from the size of the universe the sample size may depend on the following conditions.

i) The confidence limit set up for estimation;
ii) The heterogeneity of the population; and
iii) Frequency/ proportion of the trait/ attribute to be examined.

The estimation of sample size also differs according to the purpose or the parameter under investigation. For example, whether sample size is being estimated for calculating mean, or proportion, or for comparing means. For illustration, let us consider, in Box 15.2 and Box 15.3 , two cases, namely, the estimation of the mean of a normally distributed variable and the estimation of a proportion. In these cases there are two assumptions, first that sampling is simple, random and without replacement, and second, the population sampled is infinitely large.

Box 15.2 Case One

The sample size when estimating mean

It is known that the standard error of mean can be calculated from the following formula.

$SE_x = ó / vn$ 1

Where SE_x is the standard error of mean, ó is the standard deviation, and n is the sample size. Thus one can calculate sample size (n) using the following equation derived from equation 1.

$n = (ó / SE_x)^2$ 2

Sample size can be calculated using the following steps.

Step 1: One requires the standard deviation of the universe, which is unknown. A rough estimate of this measure, however, is sufficient for suggesting sample size.

a) In many instances, the experience with similar problems will be a good guide for making this estimate of the standard deviation.

b) In other instances, an exploratory sample study on a small scale may be conducted in order to arrive at an estimate of ó.

c) To estimate the standard deviation of the universe, the range of the values in the universe may be estimated and used as a guide. It is known that in normal distribution the range is about six times the standard deviation. For practical purposes, an estimate of somewhere around one-fifth of the estimated range is often used.

Suppose the range is roughly 300; that is, the difference between the lowest value in the universe and its highest value is 300. One-fifth of this rough estimate is 60. Therefore, one may take 60 as a rough approximation of ó.

Step 2: It must be decided how precise one wants the future sampling estimate to be. Thus, one may state that the estimate of the true mean is sufficiently precise if confidence limits of 12 are attached to it. Such an answer might be practicable for this particular problem.

Step 3: In this step the researcher has to decide the confidence limit. He may wish to be almost certain or be satisfied with, say, a 95% degree of confidence, that the specified limits will contain the true mean. The degree of confidence decided upon makes it possible to translate the interval decided upon in step 2 into standard error. If one is to be practically certain that true mean will lie within the interval of ±12 around the sample mean then the interval of ±12 becomes ± 3 SE_x. Therefore, $SE_x = 4$. If, on the other hand, one is willing to settle for a 95% degree of confidence, then ±12 becomes ± 2 SE_x and $SE_x = 6$.

Using equation 2, the sample size for the above example will be

1) **Case 1-**
At the level of practical certainty:
Sample size (n) = $(60 / 4)^2 = 15^2 = 225$

2) **Case 2-**
At the level of 95% confidence limit:
Sample size (n) = $(60 / 6)^2 = 10^2 = 100$
(In case 1 $SE_x = 4$, whereas in case 2 $SE_x = 6$)

Thus, in the above example, the sample size should be somewhere around 225 if one wishes to be practically certain that true mean will lie within an interval of ±12; but the sample need contain only 100

items if one settles for a 95% degree of confidence that true mean will lie within an interval of ±12.

Sometimes the acceptable difference between the sample and its true mean is expressed in percentage (say 3%) rather than absolute (as for example, ±12 in step 2 of the above example). Suppose the expected mean is around 500 then the acceptable interval would be ±15. But this necessitates an approximate knowledge of the expected mean.

Box 15.3 Case Two

Sample size when sampling for proportion

Consider the estimation of the proportion of individuals in a population with some particular attribute, for example those who own tractors for agriculture. This proportion, though not precisely known to the investigator, is generally known to him to an order of magnitude at least; that is to say, he will often know that owning tractors is quite rare (say, less than 3 in 1,000 persons), somewhat infrequent (3 in 100 to 3 in 1,000 persons), fairly common (3 in 10 to 3 in 100), or very common (more than 3 in 10). If owning tractors is known to be more infrequent than 3 in 100, a simple random sampling would invariably be much too inefficient and the other sampling methods appropriate to the estimation of rare events should be used. To assume random sampling amounts to assuming that the investigator's interest centers on only those attributes whose frequencies are at least 3 in 100. Even within these limits it is clear that if the population proportion is to be known exactly, the entire population must be examined. This is impracticable and generally unnecessary, for the investigator usually does not require this degree of exactness. His requirements are related, of course, to the use to which the estimate (or estimates) is to be put, and thus may vary from one investigator to another and with the proportion itself.

It is known that the standard error of a proportion can be calculated from the following formula.

$SE_p = v\ (PQ / n)$ 3

$SE_p^2 = PQ / n$ 4

Where, SE_p is the standard error of proportion, P is the proportion of an attribute in a population and Q is = 1 - P, and n is the sample size. Thus one can calculate sample size (n) using the following equation derived from equation 4.

$n = PQ / SE_p^2$ 5

Sample size can be calculated using the following steps:

Step 1: One requires an estimate of P, from which Q follows (Q = 1 - P), which is, of course, unknown. A rough estimate of this measure, however, is sufficient for suggesting sample size.

1) In many instances, experience with similar problems will be a good guide for making this estimate of the proportion.

2) In other instances, an exploratory sample study on a small scale may be conducted in order to arrive at an estimate of proportion.

If, however, neither of these two approaches is possible, then one can conservatively assume that P = 50% which leads to a larger sample size than any other value of P. This is because in a 50% - 50% break-up, the numerator (PQ) in the formula in equation 5 ($n = PQ / SE_p^2$), is the largest. However, for the following example, let us consider that P = 30% or 0.3.

Step 2: It must be decided upon how precise one wants the sampling estimate to be. The researcher may consider an interval of, say, ± 6% around a sample proportion as satisfactory in this situation.

Step 3: In this step the researcher has to decide the confidence limit. He may wish to be almost certain or be satisfied with, say, a 95% degree of confidence, that the specified limits will contain the true mean. In the former case, ± 6% will be equal to ± 3 SE_p and consequently SE_p =± 2%, whereas in the latter case, ± 6% will be equal to ± 2 SE_p and SE_p =± 3%.

Using equation 4, the sample size for the above example will be

1) Case 1 -
At the level of practical certainty:
Sample size (n) = $(0.3*0.7) / (0.02)^2 = 0.21 / 0.0004 = 525$

2) Case 2 -
At the level of 95% confidence limit:
Sample size (n) = $(0.3*0.7) / (0.03)^2 = 0.21 / 0.0009 = 233$
(P = 30% or 0.3; in case 1 SE_p = .02, whereas in case 2 SE_p = .03)

The use of a formula to obtain an estimate of sample size does not give us more than a rough approximation. In practice it is advisable to take the sample-size estimate as a bare minimum, to be increased for safety.

Let us now complete the Reflection and Action 15.3.

Reflection and Action 15.3

Suppose in your research project you wish to estimate sample size for calculating mean and the assumption is that sampling is simple and the population sampled is infinitely large. Further, you are in the stage of taking the three steps as elaborated in Case One given in the text, the exercise for you is to work out in detail each step and write it down in the fashion given just after Box 15.2.

15.5 Conclusion

Unit 15 discussed the important subject of sampling and provided you with relevant information on different methods of sampling. Further, it brought to you the skills of calculating the sample size.

You may like to keep in mind what Mitchell (1984: 239) said about sampling theory in statistics that it "devotes itself to providing numerical

estimates of the likelihood that the population values be within some defined range of that established from the sample - provided that the sample has been chosen in such a way as to meet the mathematical conditions to justify the computation of the probabilities concerned." Further he clarified about another type of inference that is derived while using quantitative data to support theoretical interpretation and said, "The sophistication and elaboration for choosing a 'representative' sample in this restricted sense has overshadowed the other kind of inference involved when analytical statements are made from associations uncovered in a statistical sample. This is the inference that the theoretical relationship among conceptually defined elements in the sample will also apply in the parent population. The basis of an inference of this sort is the cogency of the theoretical argument linking the elements in an intelligible way rather than the statistical representativeness of the sample."

Further Reading

Burgess, R.G. (ed) 1982. *Field Research: A Sourcebook and Field Manual.* (Contemporary Social Research 4). George Allen and Unwin: London (Read page 76 onward fro discussions of random and non-random sampling)

Denizen, N.K. (ed.) 1970. *Sociological Methods: A Sourcebook.* Butterworths: London (Read page 81 onward for useful information on sampling techniques).

Unit 16

Measures of Central Tendency

Contents

Learning Objectives

It is expected that after reading Unit 16 you would be able to

- Understand the procedure of arriving at measures of central tendency of the data collected
- Work out the ways of finding out mean, mode and median measures of central tendency
- Decide which of the three measures is more appropriate in the case of your data.

16.1 Introduction

After dealing with the skills of sampling techniques for studying large complex social groups, we would now discuss the matter of measuring central tendenc and its application.

Unit 16 deals with the basic measures of central tendency and their application for those of you who may lack a strong background in mathematics. In doing so, complex mathematical derivations of formulae have been omitted. Besides a minimal number of essential 'shorthand' mathematical symbols, and familiar examples drawn from social science data are presented in a non-mathematical form.

16.2 Mean

Mean[@] is the most common and widely used measure of central tendency. Each observation in a population may be referred to as X_i (read "X sub i") value. Thus, one observation might be denoted as X_1, another as X_2, a third as X_3, and so on. The subscript i might be any integer value up through *N*, the total number of X_i values in the population. The mean of the population is denoted by the Greek letter μ (lower case mu).

Calculating the mean from ungrouped data

Mean *(M)* is the most familiar and useful measure used to describe the central tendency average of a distribution of scores for any group of individuals, objects or events. It is computed by dividing the sum of the

scores by the total number of scores.

$M = \Sigma X_i / N$ 1

Where, M is the mean (sample), X_i are the scores, *N* is the total number of scores and Σ is 'the sum of'. See Box 16.1 and Box 16.2 for examples 1 and 2.

Box 16.1

Example 1: The Number of Cattle Owned by Members of a Community is Recorded Below.

12, 11, 13, 20, 16, 18, 19, 17, 22 and 23

$\Sigma X_i = 12 + 11 + 13 + 20 + 16 + 18 + 19 + 17 + 22 + 23 = 170$

$N = 10$

$M = \Sigma X_i / N$; $M = 170 / 10 = 17$

The mean is the balance point in a distribution such that if you subtract each value in the distribution from the mean and add all these deviation scores, the result will be zero.

Calculating mean from grouped data

Calculation of mean from grouped data is slightly different from calculation from ungrouped data.

$M = \Sigma F_i * X_i / \Sigma F_i$ 2

where, M is the mean, X_i are the midpoint of class intervals, F_i are the number of cases in various intervals, ΣF_i is the total number of scores or sum of frequencies of various intervals.

Box 16.2

Example 2: Following is the frequency (8, 9, 12, 9, 7, and 5) of households in a community owning numbers of chickens, arranged in six groups (1-3, 4-6, 7-9, 10-12, 13-16 and 16-18).

Number of Chickens	Mid-Point of the Interval (X_i)	Frequency: Number of Households (F_i)	$F_i * X_i$
1 -3	2	8	16
4 - 6	5	9	45
7 - 9	8	12	96
10 - 12	11	9	99
13 - 16	14	7	98
16 - 18	17	5	22
		50	376

$\Sigma F_i * X_i = 376$ $\Sigma F_i = 50$

$M = \Sigma F_i * X_i / \Sigma F_i = 376 / 50 = 7.52$

A short method of calculating mean from grouped data

There is a shorter way of calculating mean from grouped data, which saves time and labour in computation, particularly when one has to deal with a large number of cases. It involves the assumption of mean and making a guess at identifying the interval in which the mean probably falls (generally among the central groups of intervals). A different guess of the interval alters calculations, but not the mean.

$$\text{Mean } (M) = AM + ((\Sigma F_i * D_i \,/\, \Sigma F_i)) * i \qquad \ldots\ldots\ldots 3$$

Also,

$$D_i = (AM - X_i) \,/\, i \qquad \ldots\ldots\ldots 4$$

Where, M is the mean, AM = Assumed mean, X_i are the midpoint of class intervals, F_i are the number of cases in various intervals, Σ is the symbol of sum total, D_i are the deviations of the midpoints of the various classes from the midpoints of the class having the assumed mean divided by the size of the class interval (equation 4) and i is the size of the class intervals. See Box 16.3 for example 3.

Box 16.3 Example 3: Marital Distance (the distance between the villages of the spouse)

The marital distance was investigated in a community. Following was the frequency (88, 93, 72, 97, 79, and 54) when the data were arranged in six groups according to marital distance (25 - 30, 30 - 35, 35 - 40, 40 - 45, 45 - 50, 50 -55). Let us find the mean marital distance.

Marital Distance (km)	Frequency (F_i):	Mid-Point of the Interval (X_i)	$D_i = (AM - X_i)/ i$	$F_i * D_i$
25 -30	88	27.5	+3	264
30 -35	93	32.5	+2	186
35 -40	72	37.5	+1	72
40 -45	97	AM=42.5	0	0
45 -50	79	47.5	-1	-79
50 -55	54	52.5	-2	-108
	483			335

$AM = 42.5 \qquad \Sigma F_i = 483 \qquad \Sigma F_i * D_i = 335 \qquad i = 5\ (30\text{-}25)$

$$\text{Mean } (M) = AM + ((\Sigma F_i * D_i \,/\, \Sigma F_i)) * i$$

$$= 42.5 + (335/\ 483) * 5 = 42.5 + 3.468 = 45.968$$

After the three examples for calculating mean for ungrouped and grouped data, we would now discuss the technique of finding the Median.

16.3 Median

Median[@] is the score that divides the distribution into halves; half of the scores are above the median and the other half are below it when the data are arranged in a numerical order. Median is also referred to as the score at the 50th percentile in the distribution.

Calculating median from ungrouped data

- ❖ Arrange the series in numerical order (ascending or descending).
- ❖ Find the median location of *N* numbers by the formula (*N* + 1) / 2. When N is an odd number, for example 7 then the value of the 4th item ((7+1)/2 = 4) is the median. For example in the following ordered distribution the value of 4th item, i.e. 9 is the median.

  2, 5, 8, 9, 12, 16, 16
- ❖ Whereas, when N is an even number, say 12 then the median is half-way between the 6th and 7th items ((12+1)/2 = 6.5).

See Box 16.4 for examples 4 and 5.

Box 16.4 Example 4: Finding the Median

When N is an odd number: Find the median in the distribution of numbers: 1, 13, 8, 3, 4, 11, and 7.
The median location is (*N* + 1) / 2 or (7 + 1) / 2 = 4.
The ordered distribution is: 1, 3, 4, 7, 8, 11 and 13.
The value of 4th item in the distribution is 7 and thus median is 7.

Example 5: ***When N is an even number:***

Find the median in the distribution of numbers: 1, 8, 3, 13, 11, and 7.
The median location is (6 + 1) / 2 = 3.5.
The ordered distribution is 1, 3, 7, 8, 11 and 13.
The halfway value between the 3rd and 4th item is 7.5 ((7+8) / 2), and thus median is 7.5.

Calculating median from grouped data

Finding the median score in the frequency distribution below involves five steps.

Step 1: Divide the total number (N or ΣF_i) by two.

Step 2: Start at the low end of the frequency distribution and sum the scores in each interval until the interval containing the median is reached (C. F.).

Step 3: Subtract the sum obtained in step two above from the number necessary (calculated at step 1) to reach the median (N/2 - C. F.).

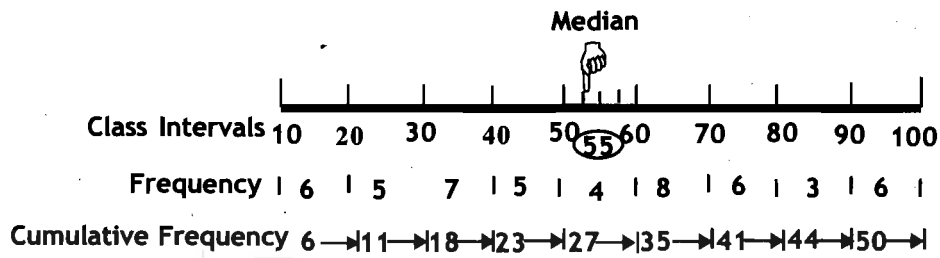
Step 4: Now calculate the proportion of the median interval that must be added to its lower limit in order to reach the median score. This is done by dividing the number obtained in step 3 above by the number of scores (f) in the median interval and then multiplying by the size of the class interval (i), i.e. [(N /2 - C.F.) / f] *i.

Step 5: Finally, add the number obtained in step 4 above to the exact lower limit of the median interval.

$$\text{Median} = L + [(N/2 - \text{C.F.}) / f] * i$$

Where, L = the exact lower limit of the median interval, N = the total number of scores; C.F. = the sum of the scores in the intervals below the median interval, f = the number of scores in the median interval; i = the size of the class interval.

Graphical representation of calculating the median from grouped data



See Box 16.5 for example 6 for finding the media for grouped data.

Box 16.5 Example 6: Find the Median of the following distribution

Class Interval	18-21	21-24	24-27	27-30	30-33	33-36	36-39	39-42	42-45	45-48	48-51
Frequency											

Class Interval	Frequency	Cumulative Frequency
18-21	1	1
21-24	2	3
24-27	3	6
27-30	6	12
30-33	7	19
33-36	8	27
36-39	8	35
39-42	6	41
42-45	4	45
45-48	3	48
48-51	2	50
Total	$\Sigma F_i = 50$	

$$\text{Median} = L + [(N/2 - \text{C.F.}) / f] * i$$

N or $?F_i / 2 = 50/2 = 25$

Lower limit of the median class (L) = 33

Cumulative frequency of the class preceding the median class (C.F.) = 19

Cumulative frequency of the class preceding the median class (C.F.) = 19
Frequency of the median class (f) = 8
Size of the class interval = 3
Median = 33 + [(25-19) / 8)] *3 = 33 + 2.25 = 35.25
Let us now complete Reflection and Action 16.1 for checking if the calculation methods have now become clearer and easier to perform.

After the Reflection and Action 16.1, you would learn about calculating mode from ungrouped and grouped data.

Reflection and Action 16.1

Following the examples given in the text for calculating the mean and median for ungrouped and grouped data and the short method of calculating mean of grouped data, provide your own examples of each of the five calculations in the manner similar to examples in the text. This exercise would provide you an opportunity of practicing such calculations. These calculation exercises would come in handy while you would carry out your own mini research project.

16.4 Mode

Mode[◎] of a distribution is simply defined as the most frequent or common score in the distribution. Mode is the point (or value) of X that corresponds to the highest point on the distribution. If the highest frequency is shared by more than one value, the distribution is said to be **multimodal**. It is not uncommon to see distributions that are bimodal reflecting peaks in scoring at two different points in the distribution.

Calculating mode from ungrouped data

The most frequent data in the series is the mode. It can be determined by viewing the series (if the series is small) or looking at the frequency distribution (if the series is large). See Box 16.6 for example 7.

Box 16.6 Example 7: Find the Mode of the following Distribution.

Serial number of family	1	2	3	4	5	6	7	8	9	10
Number of Children	1	2	3	4	3	3	2	1	2	3

In the above example 3 occurs the maximum number of times (4 times), and hence 3 is the mode of the distribution.

Calculating mode from grouped data

Mode of the grouped data can be calculated using the following steps:

Step 1: Identify the modal class (class with maximum frequency) by inspection or analysis.

Step 2: Apply the following formula

Mode = $L + [(f_m - f_1) / (f_m - f_1) + (f_m - f_2)] * i$

Or

Mode = $L + [(f_m - f_1) / (2f_m - f_1 - f_2)] * i$

Where, L = the exact lower limit of the modal interval, f_m = frequency of the modal class, f_1 = frequency of the class preceding modal class, f_2 = frequency of the class succeeding modal class, i = the size of the class interval.

You can find the graphical representation of mode in grouped data in Figure 16.1.

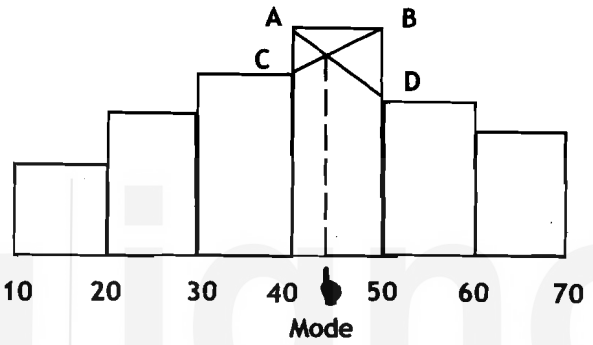


Figure 16.1 Graphical Representation of Mode in Grouped Data

The sample mode is the best estimate of population mode. When one samples a symmetrical unimodal population, mode is an unbiased and consistent estimate of mean and median, but it is relatively inefficient and should not be so used. As a measure of central tendency, mode is affected by skewness less than is mean or median, but it is affected by sampling more than these other two measures. Mode, but neither median nor mean, may be used for data on nominal, as well as the ordinal, interval, and ratio scales of measurement. Mode is not used often in social or biological researches, although it is often interesting to report the number of modes detected in a population, if there are more than one. See Box 16.7 for example 8.

Box 16.7 Example 8: Find the Modal Income on the Basis of the Following Data.

Income (in Thousands)	5 - 10	10 - 16	16 - 20	20 - 25	25 - 30	30 -35
No. of Households	8	16	29	22	14	12

Income (in Thousands)	No. of Households
5 - 10	8
10 - 15	15 f_1
Modal Class 15 - 20	29 f_m
20 - 25	22 f_2
25 - 30	14
30 - 35	12

Mode lies in the (16 - 20) having the maximum frequency (29)
Lower limit of the modal class = 16
Frequency of the modal class (f_m) = **29**
Frequency of the class preceding modal class (f_1) = 16
Frequency of the class succeeding modal class (f_2) = 22
Size of the class interval = 5
Mode = $L + [(f_m - f_1) / (2f_m - f_1 - f_2)] * i$
Mode = 16 + [(29 - 16) / (2*29 - 16 - 22)] *5 = 16 + (14 / 21) * 5 = 16 + 3.33 = 18.33
The modal income is 18.33 thousands.

After learnign about mean, median and mode, we will discuss in Section 16.5 the relationship among the three measures of central tendency. But before going on to Section 16.5, let us complete Reflection and Action 16.2.

Reflection and Action 16.2

Make a graphical representation of mode in grouped data of your choice along the lines of Figure 16.1. You may then use similar type of graphic representation of grouped data in your own mini research project.

16.5 Relationship between mean, mode and median

Mean, mode and median (the three measures of central tendency) are related to each other and can be calculated using the following equation.

Mode = 3 * Median - 2 * Mean

The values of mean, mode and median are the same when the frequency is normally distributed, but their values differ when the frequency is positively or negatively skewed.

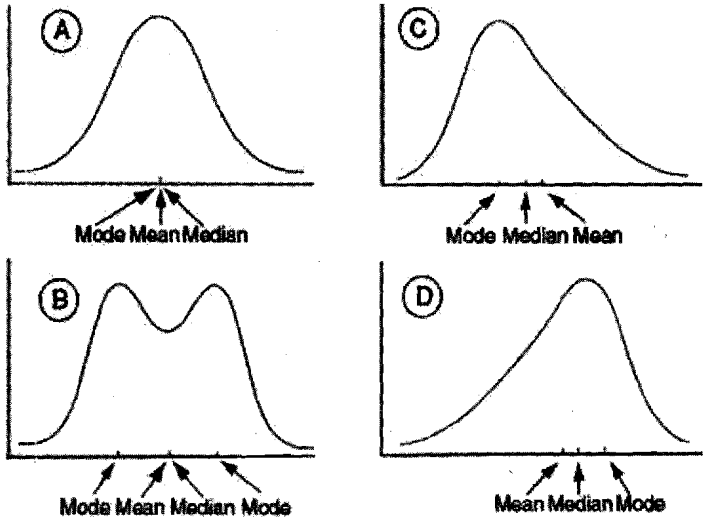


Fig. 16.2: Relationship of Mean, Mode and Median in various Types of Frequency Distributions

Fig. 16.2 shows the relationship of mean, mode and median in various types of frequency distributions: (A) Normal distribution (B) Bimodal distribution (C) Positively skewed distribution (D) Negatively skewed distribution. Values of the variables are along x axis and the frequencies are along y axis.

After learning about the relationship among the three measure of central tendency, let us find out how to decide which of the three to choose for one's research.

16.6 Choosing a measure of central tendency

Sometimes the researcher has to decide which of the three measures of central tendency to use. The following advice may be of help.

Mean is doubtless the most commonly used measure of central tendency. It is the only one of the three measures which uses all the information available in a set of data, that is to say, it reflects the value of each score in a distribution. It has the decided advantage of being capable of combining with the means of other groups measured on the same variable. For example, from the average unemployment levels in various states of India one can compute the overall mean unemployment rate of India. Since neither the median nor the mode is based on arithmetic, this useful application is not possible. The precisely defined mathematical value of the mean allows the other advanced statistical techniques to be based on it too.

There are occasions, however, when taking into account the value of every score in a distribution can give a distorted picture of the data. For example, marriage distance (the distance between the places of residence of the two partners) in five cases is 40, 60, 60, 80 and 810. Without the very atypical score of 810, the mean score of the group is 60 and the median, likewise, is 60. The effect of introducing the score of 810 is to pull the mean in the direction of that extreme value. The mean now becomes 210, a value that is unrepresentative of the series. The median remains 60, providing a more realistic description of the distribution than the mean.

With these observations in mind:

Use the mean

i) When the scores in a distribution are more or less symmetrically grouped about a central point.

ii) When the research problem requires a measure of central tendency that will also form the basis of other statistics (such as measures of variability or measures of association).

iii) When the research problem requires the combination of mean with the means of other groups measured on the same variable.

iv) To measure the central tendency in a sample of observations when one needs to estimate the value of a corresponding mean

of the population from which the sample is taken.

v) When the interval level or ratio level data providing that the distribution of scores approximates a normal curve.

Use the median

i) When the research problem calls for knowledge of the exact midpoint of a distribution.

ii) When extreme scores are there in the series, as they distort the mean, but not the median. Particularly, when dealing with 'oddly-shaped' distributions, for example, those in which a high proportion of extremely high scores occur as well as a low proportion of extremely low ones.

Use the mode

i) When all that is required is a quick and appropriate way of determining central tendency.

ii) When in referring to what is 'average', the word is used in the sense of the 'typical' or the 'most usual'. For example, in talking about the average take-home pay of the coffee plantation worker, it is the modal wage that is being alluded to rather than an exact arithmetic average.

Reflection and Action 16.3

Provide examples of data that require mean, median and mode type of calculations for reflecting the central tendency of the data.

16.7 Conclusion

Succinctly, mode would be the appropriate statistic to use as a measure of the 'most fashionable' or 'most popular' when data are collected using a nominal scale. *Median* would generally be associated with the ordinal level data. *Mean* will be used with interval level or ratio level data providing that the distribution of scores approximates a normal curve.

You can take mean to be a mathematical measure and median mode to be the positional measures. You can always cluster your observations around a central value. A central value manifests both the distribution and the comparison of various distributions. It is always useful for a researcher to provide measures that indicate the average feature of a frequency distribution. Unit 16 has discussed the three measure of central tendency and provided skills of basic statistical tools for application in your research.

It would have become apparent to you that the three measures of the central tendency, namely, i) average of all the values in the distribution or mean, ii) mid-point of the distribution or median and iii) highest

density in the distribution or mode, are not to applied in a mechanical way. In the light of the objective of your study you would need to determine when you are to use which measure. You have learnt in Unit 16 that a graphic representation of distributions shows either a symmetrical or a skewed pattern. In symmetrical type, you will find that the three values coincide. This provides you the option of using the mean. In the case of bi-modal or multi-modal representation, you would do better to use the mode. In skewed distribution, if the tail is on the right side, it indicates the positive skewing of distribution. If the tail is on the left side, it shows the negative skewing of distribution. For both the negative and the positive types of skewing of distribution, you would do better by using the median measure of central tendency. You may want to work out with the help of Unit 16 the type of measure of central tendency you will use in your mini research project.

Further Reading

Black, Thomas R. 1999. *Doing Quantitative Research in the Social Science*. An Integrated Approach to Research Design, Measurement and Statistics

Nachmias, David and Chava Nachmias 1981. *Research Methods in Social Sciences*. St. Martin Press: New York.

Unit 17

Measures of Dispersion and Variability

Contents

Learning Objectives

It is expected that after reading Unit 17 you would be able to

- ❖ Obtain a measure of dispersion of data
- ❖ Explain the meaning of the term 'range' and work out how to measure the range of one's data
- ❖ Discuss the element of variance in one's data and find out the standard variation in it
- ❖ Work out the coefficient of variation in the data.

17.1 Introduction

In addition to a measure of central tendency, it is generally desirable to have a measure of dispersion of data. A measure of dispersion (or a measure of variability©, as it is sometimes called) is an indication of the clustering of measurements around the center of the distribution, or, conversely, it is an indication of how variable the measurements are. Sanders (1955) held that you need to measure dispersion to evaluate the extent to which the average value depicts the data. Another reason for measuring dispersion is to find out the spread in order to improve or control the existing variations.

17.2 The Range

The difference between the highest and the lowest measurements in a group of data is termed range©. If sample measurements are arranged in an increasing order of magnitude, as if the median were about to be determined, then

$$\text{Sample range} = X_n - X_1 \qquad \text{.....1}$$

Where, X_1 and X_n are the lowest and the highest value of the series respectively.

See Box 17.1 for Example 1.

Box 17.1 Example 1

The number of cattle owned by members of a community is recorded as: 12, 11, 13, 20, 15, 18, 19, 17, 22 and 23. Calculate the range.

$X_1 = 11$; $X_n = 23$

Sample range = 23 - 11 = 12

The range is a relatively crude measure of dispersion, inasmuch as it does not take into account any measurement except the highest and the lowest. Furthermore since it is unlikely that a sample will contain both the highest and the lowest values in the population, the sample range usually underestimates the population range; therefore, it is a biased and inefficient estimator. Nonetheless, it is useful in some circumstances to present the sample range as an estimate (although a poor one) of the population range. Whenever the range is specified in reporting data, it is usually a good practice to report another measure of dispersion as well.

The Mean Deviation

It is clear that no information is provided by the range about the distribution of the measurements in the middle. Since the mean is so useful a measure of central tendency, one might express dispersion in terms of deviations from the mean.

The sum of all deviations from the mean ($(\Sigma(X_i - M))$ will always be zero, therefore such a summation would be useless as a measure of dispersion. On the other hand, the sum of the absolute values of the deviation from the mean expresses dispersion about the mean. Dividing this sum by the total number yields a measure that is known as *mean deviation*, or *mean absolute deviation* of the sample, is obtained.

$$\text{Sample mean deviation} = (\Sigma | X_i - M |) / n \quad \text{.....2}$$

Where, M is the mean (sample), X_i are the scores, n is the total number of scores and ?is 'the sum of' and the vertical lines indicate that the values are absolute (irrespective of sign). See Box 17.2 for example 2.

Box 17.2 Example 2

The number of cattle owned by members of a community is recorded as: 12, 11, 13, 20, 15, 18, 19, 17, 22 and 23. Calculate the mean deviation.

$\Sigma X_i = 12 + 11 + 13 + 20 + 15 + 18 + 19 + 17 + 22 + 23 = 170$

$N = 10$

$M = \Sigma X_i / N$; $M = 170 / 10 = 17$

$(\Sigma | X_i - M |) = (12 - 17) + (11 - 17) + (13 - 17) + (20 - 17) + (15 - 17) + (18 - 17) + (19 - 17) + (17 - 17) + (22 - 17) + (23 - 17) = 5 + 6 + 4 + 3 + 2 + 1 + 2 + 0 + 5 + 6 = 34$

Sample mean deviation = 34 / 10 = 3.4

It is possible that the two samples may have the same range, but not the mean deviation. Mean deviation can also be defined by using the sum of the absolute deviations from the median rather than from the mean.

17.3 The Variance

Another method of eliminating the signs of deviations from the mean is to square the deviations. The sum of the square of deviation from the mean is called the *sum of squares*, abbreviated SS, and is defined as follows:

$$\text{Sample SS} = \Sigma (X_i - M)^2 \quad \text{.....3}$$

Where, M is the mean (sample), X_i are the scores, and Σis 'the sum of'.

From the sample SS, population SS can be estimated.

$$\text{Population SS} = \Sigma (X_i - \mu)^2 \quad \text{.....4}$$

Where M is the mean (sample), X_i are the scores, and ?is 'the sum of'.

The mean sum of square is called *variance* (or *mean square*, the latter being short for *mean squared deviation*), and for a population is denoted by ó ² ("sigma squared", using the lowercase Greek letter).

Calculating variance from ungrouped data

$$\text{Population Variance} = ó^2 = \Sigma(X_i - \mu)^2 / N \quad \text{.....5}$$

The best estimate of the population variance, ó ², is the sample variance, s^2:

$$\text{Sample Variance} = s^2 = \Sigma(X_i - M)^2 / (n - 1) \quad \text{.....6}$$

Where M is the mean (sample), X_i are the scores, n is the total number of scores (sample) and Σ is 'the sum of'.

The replacement of μ by M and N by n in the above equation results in a quantity which is a biased estimate of ó ². Dividing the sample's sum of squares by n-1 (called the degree of freedom, abbreviated DF) rather than by n, yields an unbiased estimate and the above equation should be used to calculate the sample variance. If all observations are equal, then there is no variability and $s^2 = 0$; and s^2 becomes increasingly large as the amount of variability, or dispersion, increases. Since s^2 is a mean sum of squares, it can never be a negative quantity.

The variance expresses the same type of information as does the mean deviation, but it has certain important properties relative to probability and hypothesis testing that makes it distinctly superior. Thus, the mean deviation is very seldom encountered in social or bio-statistical analysis.

The variance has square units. If measurements are in grams, their variance will be in grams squared, or if the measurements are in cubic centimeters, their variance will be in terms of cubic centimeters squared, even though such squared units have no physical interpretation.

The sample variance© can be calculated using the following formula

$$\text{Sample variance} = s^2 = ((\Sigma X_i^2) - (\Sigma X_i)^2 / n)) / (n - 1) \quad \text{.....7}$$

The above formula is often called the machine formula, because of its computational advantages. There are, in fact, two major advantages in

calculating SS by Equation 7 rather than by Equation 6. First, here fewer computational steps are involved, a fact that decreases the chance of error. On a good desk calculator, the summed quantities, ΣX_i and ΣX_i^2 can both be obtained with only one pass through the data, whereas Equation 6 requires one pass through the data to calculate M, and at least one more pass to calculate and sum the squares of the deviations, $X_i - M$. Second, there may be a good deal of rounding error in calculating each $X_i - M$, a situation which leads to decreased accuracy in computation, but which is avoided by the use of Equation 7. See Box 17.3 for example 3.

Box 17.3 Example 3

The number of cattle owned by members of a community is recorded as: 12, 11, 13, 20, 15, 18, 19, 17, 22 and 23. Calculate the sample variance.

$\Sigma X_i = 12 + 11 + 13 + 20 + 15 + 18 + 19 + 17 + 22 + 23 = 170$

$n = 10$

$M = \Sigma X_i / n$; $M = 170 / 10 = 17$

X_i	12	11	13	20	15	18	19	17	22	23	$\Sigma X_i = 170$
$X_i - M$	-5	-6	-4	+3	-2	+1	+2	0	+5	+6	
$(X_i - M)^2$	25	36	17	9	4	1	4	0	25	36	$\Sigma(X_i - M)^2 = 156$

Sample variance = $s^2 = \Sigma (X_i - M)^2 / (n - 1) = 156 / 9 = 17.33$

Alternate formula (often called machine formula)

Sample variance = $s^2 = ((\Sigma X_i^2) - (\Sigma X_i)^2 / n)) / (n - 1)$

X_i	12	11	13	20	15	18	19	17	22	23	$\Sigma X_i = 170$
X_i^2	144	121	179	400	225	324	361	289	484	529	$\Sigma X_i^2 = 3046$

Sample variance = $s^2 = (3046 - ((170)^2)/ 10) / 9 = 156 / 9 = 17.33$

Calculating the variance from grouped data

The sample variance in the grouped data can be calculated using the following formula.

$$\text{Sample Variance} = s^2 = \Sigma f_i (X_i - M)^2 / (n - 1) \quad \text{.....8}$$

Where, M is the mean (sample), f is the frequency of observations with magnitude X_i, n is the total number of scores (sample) and Σis 'the sum of'.

The manual calculation becomes complex, if the mean value is having several places after decimal. A commonly used method is from assumed mean. The formula is listed below.

$$\text{Sample Variance} = s^2 = \{(\Sigma f_i * d_i^2)/ n - (\Sigma f_i * d_i / n)^2\} * i \quad \text{.....9}$$

$$d_i = (X_i - A)/ i$$

Where, i is the size of the class interval, f_i is the frequency of observations with magnitude X_i, n is the total number of scores (sample) and Σis 'the sum of'. See Box 17.4 for example 4.

Box 17.4 Example 4

Agricultural land (in acres) owned by various households is grouped under the following seven groups. The frequency of households in each category is listed below. Find the variance in the land owning.

Land Owned (in Acres)	Frequency (F_i)	Mid-Point of the Interval (X_i)	$d_i = (X_i - A) / i$	$f_i * d_i$	d_i^2	$f_i * d_i^2$
20 - 30	18	25	-3	-54	9	172
30 - 40	19	35	-2	-38	4	76
40 - 50	12	45	-1	-12	1	2
50 - 60	19	Assumed Mean (A) = 55	0	0	0	0
60 - 70	17	65	1	17	1	17
70 - 80	15	75	2	30	4	60
80 - 90	10	85	3	30	9	90
	110			-27		417

$\Sigma f_i * d_i^2 = 417$ $\Sigma f_i * d_i = -27$ $n = 110$

Sample Variance = $\{(\Sigma f_i * d_i^2)/ n - (\Sigma f_i * d_i / n)^2\} * i$

Sample Variance = $\{(417/ 110) - (-27/ 110)^2\} * 10 = (3.79 - .06) / 10 = 37.3$

The variance in the grouped data can also be calculated using the following equation (often called machine formula).

Sample variance $(s^2) = ((\Sigma f_i * X_i^2) - (\Sigma f_i * X_i)^2 / n)) / (n - 1)$ 10

Where f_i is the frequency of observations with magnitude X_i.

But with a desk calculator it is often faster to use Equation 7 for each individual observation, disregarding the class groupings. See Box 17.5 for example 5.

Box 17.5 Example 5

An investigation in a community on the bride price yielded the following data. Find the variance in bride price.

Bride Price (in Thousand Rs)	Frequency (F_i)	Mid-Point of the Interval (X_i)	$f_i * X_i$	X_i^2	$f_i * X_i^2$
10 - 20	8	15	120	225	1800
20 - 30	9	25	225	625	5625
30 - 40	12	35	420	1225	14700
40 - 50	9	45	405	2025	18225
50 - 60	7	55	385	3025	21175
60 - 70	5	65	325	4225	21125
	50		1880		82650

$\Sigma f_i * X_i^2 = 82650 \quad \Sigma f_i * X_i = 1880 \quad (\Sigma f_i * X_i)^2 = (1880)^2 = 3534400$

$n = 50$

Sample variance $(s^2) = ((\Sigma f_i * X_i^2) - (\Sigma f_i * X_i)^2 / n)) / (n - 1)$

Sample variance $(s^2) = (82650 - (3534400 / 50) / 49 = (82650 - 70688) / 49 = 11962 / 49 = 244.12$

Reflection and Action 17.1

Following the examples in the text, provide your own examples for calculating variance from ungrouped and grouped data.

17.4 The Standard Deviation

The standard deviation© is the positive square root of the variance; therefore, it has the same units as the original measurements. It can be calculated using the following formula.

Standard deviation (s) = v (Sample Variance)

In Example 5, you found the sample variance to be = 244.12, and therefore you can work out the standard deviation to be (s) = v 244.12 = 15.62

Thus various examples given above for the calculation of variance explain the procedure of calculating standard deviation.

17.5 Coefficient of Variation

Ratio scales are useful in social science research when an investigator is interested in the variability of a sample on one characteristic as compared to another.

The coefficient of variation is the percentage ratio of standard deviation to mean and it is calculated using the following formula.

Coefficient of variation = standard deviation *100 / Mean

It is a useful measure of dispersion, when comparison of variability is being made between the variables of unequal magnitude and/ or have different units of measurements, for example, height and weight.

In example 4, you would find that

Mean (M) = AM + $(\Sigma f_i * d_{ij} / n) * i$ = 55 + (-27 / 110) *10 = 55 - 2.45 = 52.55 and

Standard deviation (s) = v (Sample Variance) = v37.3 = 6.107

Coefficient of variation = s *100 / M = 6.107 / 52.55 = 11.62

Reflection and Action 17.2

Work out standard deviation and coefficient of variation of the examples you selected in Reflection and action 17.1.

17.6 Conclusion

After working out in Unit 16 how to measure the central tendency in one's data, in Unit 17 you acquired the skill of measuring dispersion of data, which indicates the clustering of measurements around the center of the distribution, or, you may say that it is an indication of how variable the measurements are.

You may agree with Sanders (1955: 90-91) who said that the range is an easy measure to work out and understand because it requires only one subtraction and it places stress on the extreme values. The mean absolute deviation, on the other hand, places equal weight to the deviation in every observation and it is equally easy to work out and understand. The squaring of deviations in calculating standard deviation emphasies the extreme value. The standard deviation is a more common measure of dispersion. The value of every observation in a series affects the value of this measure. A change in the value of any observation will generate a change in the standard deviation value. Relatively few extreme values can distort its value. The standard deviation is not possible to compute from an open ended distribution. Finally, the co-efficient of deviation is similar to the range as it is based on only two values, which identify the range of the middle fifty percent of the values.It is mostly used in the sets of skewed data and it is possible to compute it in an open-ended distribution.

Further Reading

Sanders, Donald 1955, *Statistics*. McGraw-Hill: New York

Unit 18

Statistical Inference: Tests of Hypothesis

Contents

Learning Objectives

It is expected that after reading Unit 18 you would be able to

- ❖ Draw statistical inferences on the basis of the concept of probability
- ❖ Use the tool of statistical inference to test hypotheses
- ❖ Apply the tool of statistical inference for estimating the unknown parameter of the population under research.

18.1 Introduction

Unit 18 deals with statistical inference, which uses the concepts of probability to explain the element of uncertainty in decision-making. You would find that though it occupies a lower status among statistical tests, you would be able to use chi-square test in a wide variety of researches. If you have a relatively smaller sample, it would be better to use student's test that is a parametric test. You would learn in Unit 18 in detail about both the chi-square and student's tests. For hypothesis testing, Unit 18 is going to prove to be most helpful in the mini research project that you have to complete as a part of your assignment of MSO 002.

18.2 Statistical Inference

Statistical inference uses the concept of probability to deal with uncertainty in decision-making. It refers to the process of selecting and using a sample statistics to draw inferences about a population parameter, based on a sample drawn from the population. Statistical inference takes care of the two classes of problems.

A. Hypothesis testing: It tests some hypotheses about the parent population based on the sample drawn from the population.

B. Estimation: It uses the 'statistics' obtained from the sample as an estimate of the unknown 'parameter' of the population based on the sample drawn from the population.

A. Hypothesis testing

It begins with an assumption called a hypothesis that one makes about a population parameter.

Steps in testing a hypothesis

i) Formulate a hypothesis

ii) Decide an appropriate significance level

iii) Select a test criterion

iv) Carry out calculations

v) Make Decisions

Let us discuss in brief each of the five steps.

i) Formulate a hypothesis: First of all a hypothesis is set up about a population parameter. Thereafter, sample data is collected, sample statistics calculated and the information is used to assess how far the hypothesised parameter is correct. Examining the difference between the hypothesised value and the actual value of the sample's mean tests the validity of an assumption.

Conventionally, rather than a single hypothesis two are constructed. These hypotheses are constructed in such a way that if one hypothesis is accepted the other is rejected. The two hypotheses are called:

a. Null hypothesis (designated as H_0)

b. Alternative hypothesis (designated as H_A)

In the simplest form, a null hypothesis states that there is no true difference between the sample statistics and the population parameter. It asserts that the observed difference is accidental and / or unimportant arising out of the fluctuations in sampling.

A researcher, for instance, who wishes to test whether the annual per capita income in a community is higher than Rs. 10,000/- might formulate the null and alternate hypotheses as under.

Null hypothesis (H_0): $\mu = 10,000$

Alternative hypothesis (H_A): $\mu > 10,000$

In another instance, a researcher might wish to test the mean difference between the annual per capita incomes of two groups. In this case, she might formulate the null and alternate hypotheses as under.

Null hypothesis (H_0): $\mu_1 - \mu_2 = 0$

Alternative hypothesis (H_A): $\mu_1 - \mu_2 \neq 0$

ii) Decide an appropriate significance level: The next step in testing hypotheses is to set up a suitable significance level to test the validity of H_0 as against H_A. The confidence with which a null hypothesis is adopted or rejected depends on the adopted significance level.

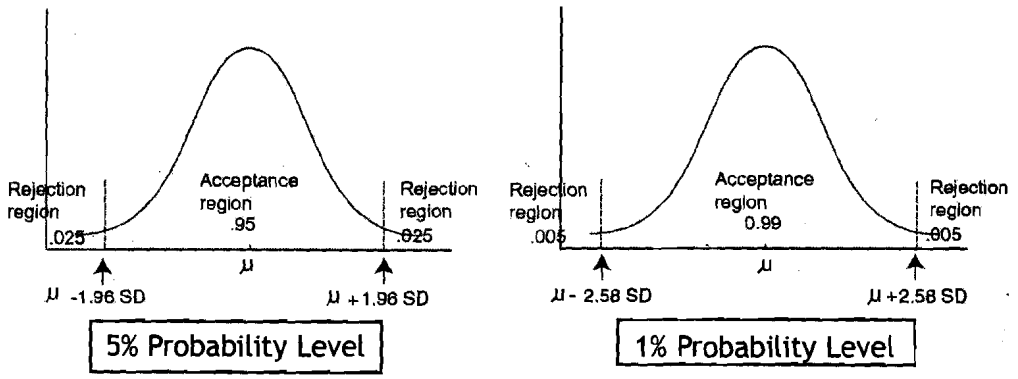


Figure18.1 Acceptance (or rejection) of Null hypothesis (two tailed) at 5% and 1%, respectively

Conventionally, the significance level is expressed as a percentage, such as 5% or 1%. In the former case, it would mean that there is 5% probability of rejecting a null hypothesis, even if true. This means that there are 5 out 100 chances that the investigator would reject a true hypothesis (see Figure 18.1).

iii) **Select a test criterion**: The next step in hypothesis testing is to set up a test criterion. An appropriate probability distribution that can be applied is selected for the particular test. Some of the common probability distributions are $\div^2$, t and F.

iv) **Carry out calculations**: Computation is carried out of various statistics and their standard errors based on sample.

B. **Make Decisions**: In this step statistical conclusions and decisions are made to reject or accept the null hypothesis, depending on whether the computed value falls in the region of acceptance or rejection (See Case 1 and Case 2 in 18.3).

Reflection and Action 18.1

Let us say that you are carrying out a research that has both the null and alternative hypotheses. You need now to set up a suitable significance level to test the validity of null hypothesis as against alternative hypothesis. For this task as well subsequent tasks, follow the procedure given in the text. Next, you would need to set up a test criterion. For this purpose select an appropriate probability distribution that can be applied for the particular test. Then carry out computation of various statistics and their standard errors. Now, based on sample statistical conclusions, make decisions to reject or accept the null hypothesis. This would depend on whether the computed value falls in the region of acceptance or rejection. Work out the steps in concrete terms of your own research project and incorporate them in your research work report.

18.3 Cases

Case 1: If the hypothesis is being tested at 5% level and the observed result has a probability of less than 5%, then the difference between the sample statistics and the population parameter is significant and cannot

be explained by chance alone. Thus the null hypothesis (or H_0) is rejected, and in turn, the alternative hypothesis (H_A) is accepted.

Case 2: If the hypothesis is being tested at 5% level and the observed result has a probability of more than 5%, then the difference between the sample statistics and the population parameter is not significant and can be explained by chance variation. Thus the null hypothesis (or H_0) is accepted, and in turn the alternative hypothesis (H_A) is rejected.

In hypothesis testing it is important to understand the following:

i) One tailed and two tailed test of hypothesis; and

ii) Type I and Type II errors

i) One-tailed and two-tailed test of hypothesis

Depending on the research problem, the null and alternate hypotheses are defined in such a way that the test is known as one-tailed or two-tailed. A two-tailed test of hypothesis will reject the null hypothesis if the sample statistic is significantly higher or lower than the population parameter. Thus, in a two-tailed test of hypothesis the rejection region is located on both the tails and the size of the rejection region is .025, whereas the central acceptance region is .95 (Fig.18 2). If the sample mean falls within μ± 1.96 SD (i.e. in the acceptance region), the hypothesis is accepted. If on the other hand, it falls beyond μ± 1.96 SD, then the hypothesis is rejected, as it will fall in the rejection region.

Let us take an example of the two-tailed hypothesis. Suppose a researcher is interested in knowing whether there is gender difference in IQ. You can formulate the following hypotheses.

IQ of Females = IQ of Males (Null hypothesis)

IQ of Females ± IQ of Males (Alternative hypothesis) or in other words, IQ of females may be lower or higher than that of males.

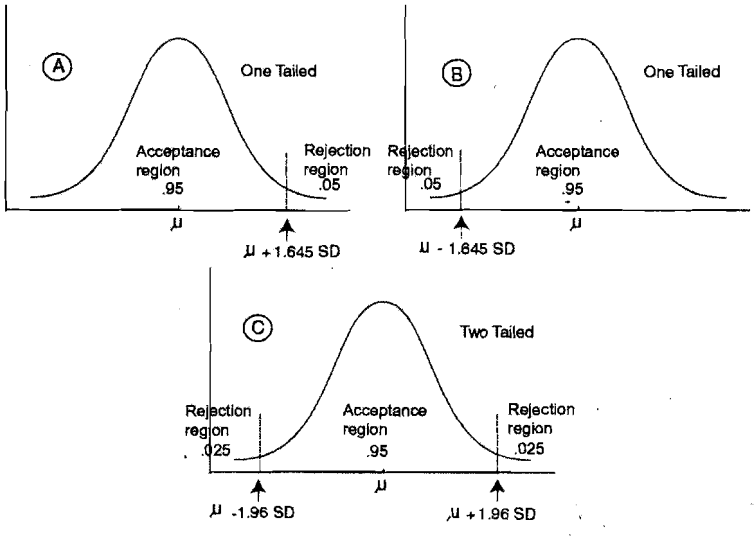


Figure 18.2 One-Tailed and Two-tailed Test of Hypothesis. (A) and (B) are One-Tailed, whereas (C) is Two-tailed.

In contrast to the two-tailed hypothesis, in one-tailed hypothesis the rejection region will be located only on one tail (see Figure 18.2). In this case, the size of the rejection region will be .05, if one is testing the hypothesis at 5% probability level. If the sample mean falls above μ + 1.645 SD (Case A: Fig. 2) or below μ - 1.645 SD (see Case B of Figure 18.2), then the hypothesis is rejected, as it will fall in the rejection region.

Let us take an example of the one-tailed hypothesis. Suppose a researcher is interested in knowing whether the IQ of females is higher than that of males. In this case, you can formulate the following hypotheses.

IQ of Females > IQ of Males (Null hypothesis)

IQ of Females = IQ of Males (Alternative hypothesis)

ii) Type I and Type II errors

i) A researcher's decision is correct when a true hypothesis is accepted and the false hypothesis is rejected. One-tailed and two-tailed test of hypothesis; and

ii) Type I and Type II errors

	Accept H_0	Reject H_0
H_0 is True	Correct Decision	Type I Error
H_0 is False	Type II Error	Correct Decision

Figure 18.3 Type I and Type II Errors in Testing a Hypothesis.

The Type I error is designated as < (alpha), whereas Type II error is designated as B (beta). It is important to note that both types of errors cannot be reduced simultaneously, as reduction in one leads to increase in the other if the sample size remains unchanged. Thus if Type I error decreases, Type II error will increase. In most of the statistical tests, the level of significance is fixed at 5% probability level (= 0.05). This means that the probability of accepting a true hypothesis is 95%. Sometimes, the level of significance is fixed at 1% probability level (=0.01). In that case, the probability of accepting a true hypothesis is 99%. In this case, accepting a false hypothesis will also increase.

Reflection and Action 18.2

Carry on with hypothesis testing with the same example as you had taken in Reflection and Action 18.1. You need now to carry out both one-tailed and two-tailed tests of your hypothesis. A two-tailed test of hypothesis will reject the null hypothesis if the sample statistic is significantly higher or lower than the population parameter. In a two-tailed test of hypothesis the rejection region is located on both the tails and the size of the rejection region is .025, whereas the central acceptance region is .95. In contrast to the two-tailed hypothesis, you would notice that in a one-tailed hypothesis the rejection region will be located only on one tail. Similarly, you would need to carry out the Type I and Type II error tests. Designate Type I error as < (alpha), and Type II error as B (beta). As pointed out in the text above, you need to remember that both types of errors cannot be reduced simultaneously, as reduction in one leads to increase in the other, if the sample size remains unchanged. If you follow the text in Section 18.3, you would be able to carry out both sets of tests on the hypothesis of your research. Make sure to include them in your research work report.

18.4 Tests of Significance

i) Chi-square test ($\div^2$)

Chi-square© is probably the most commonly used of all non-parametric tests. It is applicable when data are nominal and grouped in categories. You can examine the difference between the observed and the expected frequencies.

$$\div^2 = \Sigma ((O - E)^2 / E)$$

Where, O and E are the observed and expected frequencies respectively.

The calculated value of $\div^2$ is compared with the table value of $\div^2$ for given degrees of freedom at a certain specified level of significance (e.g. 5%). If the calculated value of $\div^2$ is higher than the table value of $\div^2$, then the difference between the theory and observation is considered to be significant. On the other hand, if the calculated value of $\div^2$ is lower than the table value of $\div^2$, then the difference between the theory and observation is considered to be non-significant.

As mentioned above, while comparing the calculated value of $\div^2$ with the table value of $\div^2$, one has to determine the degrees of freedom. Degree of freedom is the number of classes to which the values can be allocated at will or arbitrarily without defying the limitations or restrictions. For instance, if one has to choose four numbers whose sum is 100, the freedom of choice exists only for selecting three numbers, and the fourth is selected automatically. If, for example, the first three numbers are 14, 26, 32, then the fourth is fixed and must be 28 (100 - (14+ 26+ 32)). In this case the degree of freedom is three. Chi-square is used for a variety of purposes. Also there are numerous tests that are close to $\div^2$. Here, the test of goodness of fit and the test of homogeneity / association are presented.

ii) Test of goodness of fit: We often want to know whether the observed frequencies are in agreement with the probability or expected theoretical distribution or not. The following steps may be followed:

Step 1: Define null and alternative hypotheses.

Step 2: Decide probability level.

Step 3: Estimate the expected frequency E for each category based on theory and or probability.

Step 4: Calculate chi-square.

Step 5: Determine the degree of freedom.

Step 6: Compare the observed chi-square with the tabulated chi-square. Accept/ reject the null hypothesis.

See Box 18.1 for an example.

Box 18.1 Example: Test whether a form of transport is favoured more significantly than another?

	Mode of Transport					
Frequencies	*Car*	*Bus*	*Metro*	*Scooter*	*Train*	*Total*
Observed	18	21	19	20	22	100
*Expected**	20	20	20	20	20	100

Solution:

Step 1: Null hypothesis: There is no significant difference in the choice of the type of transportation.

Alternative hypothesis: There is significant difference in choice of type of transportation.

Step 2: Probability level for the hypothesis testing is 5%.

Step 3: The expected frequencies (20) in all the categories is based on the fact that there is an equal choice of the type of transportation.

Step 4: Calculations:

$\div^2 = \Sigma((O - E)^2 / E)$

$\div^2 = ((18 - 20)^2 / 20) + ((21 - 20)^2 / 20) + ((19 - 20)^2 / 20) + ((20 - 20)^2 / 20) + ((22 - 20)^2 / 20)$

$\div^2 = 4/20 + 1/20 + 1/20 + 0 + 4/20 = 10/20 = 0.5$

Step 5: Degree of freedom = k -1 = 5 -1 = 4

Step 6: The table value of chi-square at 5% probability level for 4 degree of freedom is = 9.49. The calculated value of $\div^2$ (0.5) is lower than the table value of $\div^2$ (9.49). Thus the null hypothesis is accepted and the difference between the theory and observation is non-significant and there is no significant difference in the choice of the type of transportation.

iii) Test of association/ homogeneity: This type of $\div^2$ is used for two purposes. The first purpose is to examine whether or not the two or more attributes are associated (test of association). The second purpose is to determine whether two samples are drawn from the same population or not (test of homogeneity©). In the former case the data is based on

one sample whereas in the latter, there are two or more samples.

Chi-square, a non-parametric test, is a rough estimate of confidence; it accepts weaker, less accurate data as input than the parametric tests, like t-tests and the analysis of variance, and therefore, has less status in the pantheon of statistical tests. Nonetheless, its limitations are also its strengths; because chi-square is more 'forbearing' in the data it will accept, it can be used in a wide variety of researches.

The steps in the chi-square method for the test of homogeneity remain the same as that of the test of goodness of fit, except that in step 3 the expected frequencies are calculated for each cell as illustrated.

Populations	Attribute			Total
	Category 1	Category 2	Category 3	
Population 1	A	B	C	N_1
Population 2	D	E	F	N_2
Total	N_3	N_4	N_5	N

Expected Frequency of Cell A = $(N_1 * N_3) / N$

Expected Frequency of Cell B = $(N_1 * N_4) / N$

Expected Frequency of Cell C = $(N_1 * N_5) / N$

Expected Frequency of Cell D = $(N_2 * N_3) / N$

Expected Frequency of Cell E = $(N_2 * N_4) / N$

Expected Frequency of Cell F = $(N_2 * N_5) / N$

See Box 18.2 for an example to find out if there was a difference in the income of the two groups. On the basis of this example, you may take up another case to test association. Homogeneity.

Box 18.2 Example to Examine if the Bhils and Minas Differ in their Income

Populations	Income groups		
	High	Middle	Low
Bhils	28	41	65
Minas	31	43	55

Solution:

Step 1: Null hypothesis: There is no significant difference in income between Bhils and Minas.

Alternative hypothesis: There is significant difference in income between Bhils and Minas.

Step 2: Probability level for the hypothesis testing is 5%.

Step 3: The expected frequencies are as given below.

	Income groups						Total
	Observed						
Populations	High		Middle		Low		
	Observed	Expected	Observed	Expected	Observed	Expected	
Bhils	28	*30.06*	41	*42.80*	65	*61.14*	*134*
Minas	31	*28.94*	43	*41.20*	55	*58.86*	*129*
Total	59	*59.00*	84	*84.00*	120	*120.00*	*263*

Expected Frequency of Cell A = $(N_1 * N_3) / N = 134 * 59 / 263 = 30.06$

Expected Frequency of Cell B = $(N_1 * N_4) / N = 134 * 84 / 263 = 42.80$

Expected Frequency of Cell C = $(N_1 * N_5) / N = 134 * 120 / 263 = 61.14$

Expected Frequency of Cell D = $(N_2 * N_3) / N = 129 * 59 / 263 = 28.94$

Expected Frequency of Cell E = $(N_2 * N_4) / N = 129 * 84 / 263 = 41.20$

Expected Frequency of Cell F = $(N_2 * N_5) / N = 129 * 120 / 263 = 58.86$

Step 4: Calculations:

$\div^2 = \Sigma ((O - E)^2 / E)$

$\div^2 = ((28 - 30.06)^2 / 30.06) + ((41 - 42.80)^2 / 42.80) + ((65 - 61.14)^2 / 61.14) + ((31 - 28.94)^2 / 28.94) + ((43 - 41.20)^2 / 41.20) + ((55 - 58.86)^2 / 58.86)$

$= 0.141 + 0.076 + 0.244 + 0.147 + 0.079 + 0.253 = 0.940$

Step 5: Degree of freedom = [(No. of rows -1)* (No. of column -1)] = (2-1)*(3-1) = 2

Step 6: The table value of chi-square at 5% probability level for 2 degree of freedom is = 5.991. The calculated value of $\div^2$ (0.940) is lower than the table value of $\div^2$ (5.991). Thus the null hypothesis is accepted and the difference between the theory and observation is non-significant and there is no significant difference in the income of Bhils and Minas.

There is a short cut method for the calculation of $\div^2$ if the frequency distribution is arranged in '2x2 contingency table', as illustrated in Figure 18.3.

	Variable 1 Category 1	Variable 2 Category 2	Total
Sample 1	A	B	A + B
Sample 2	C	D	C + D
Total	A + C	B + D	N = A +B +C + D

Figure 18.3 Short-cut Method to Calculate

$\div^2 = N* (A*D - B*C)^2 / (A + B) *(C + D)* (A + C)* (B + D)$

The calculated value of $\div^2$ is examined against the tabulated value at 1 d. f. at specified probability level to ascertain significance. See Box 18.3 to find out significant difference between males and females in terms of their occupations.

Box 18.3 Example to Examine Significant Gender-based Difference in Occupation as Skilled/ Unskilled Labourers

Gender	Skilled Labourers	Unskilled Labourers
Males	47	56
Females	32	71

Solution

Step 1: Null hypothesis: There is no significant sex difference in skilled and unskilled laborers.

Alternative hypothesis: There is a significant sex difference in skilled and unskilled laborers.

Step 2: Probability level for the hypothesis testing is 5%.

Step 3: Calculations:

Gender	Skilled Labourers	Unskilled Labourers	Total
Males	47	56	103
Females	32	71	103
Total	79	127	206

N=206 A*D= 3337 B*C = 1792

A + B= 103 C + D=103 A + C=79 B + D=127

$\div^2 = N^* (A^*D - B^*C)^2 / (A + B) *(C + D)^* (A + C)^* (B + D)$

$\div^2 = 206^* (3337 - 1792)^2 / 103 + 103 + 79 + 127$

$\div^2 = 491727150 / 106440097 = 4.620$

Step 4: Degree of freedom = [(No. of rows -1)* (No. of column -1)] = (2-1)*(2-1) = 1

Step 5: The table value of chi-square at 5% probability level for 1 degree of freedom is = 3.841. The calculated value of $\div^2$ (4.620) is higher than the table value of $\div^2$ (3.841).

So you can say as a conclusion that the null hypothesis is rejected and the sex difference between skilled and unskilled laborers is significant.

iv) Student's t test (t)

Student's t test is a parametric test most suitable for a small sample. It is probably the most widely used statistical test and certainly the most widely known. It is simple, straightforward, easy to use, and adaptable to a broad range of situations. No statistical toolbox should ever be without it. "Student" (real name: W. S. Gossett) developed the statistical methods to solve problems stemming from his employment in a brewery. Like chi-square, the following steps may be followed for the use of the Student's t test:

Step 1: Define null and alternative hypotheses.

Step 2: Decide probability level.

Step 3: Calculate the value of t using appropriate formula.

Step 4: Determine the degree of freedom.

Step 5: Compare the observed chi-square with the tabulated chi-square. Accept or reject the null hypothesis.

Student's t test is applied in different conditions, such as

a) To test the significance of the mean of a random sample
b) To test the difference between the means of the two independent samples
c) To test the difference between the means of the two dependent samples
d) To test the significance of the correlation coefficient.

Let us discuss each of the above conditions.

a) To test the significance of the mean of a random sample: This test is used when the researcher is interested in examining whether the mean of a sample from the normal population deviates significantly from the hypothetical population mean. The following formula is used for its calculation:

$$t = \{(M - \mu) * vn\} / S$$

When using actual mean:

$$S = v\ [\Sigma(X - M)^2 / (n - 1)]$$

When using assumed mean

$$S = v\ [\{\Sigma d^2 - (d_m)^{2*}n\} / (n - 1)]$$

Where M and μ are the means of the sample and population respectively;

n is the sample size

S is the standard deviation of the sample.

d = X - A, X being the variable

d_m is the mean of deviation

A is the assumed mean. Let us take an example, in Box 18. 4, of testing the mean nutritional intake.

Box 18.4 Example to Test The Mean Nutritional Intake in the Population with 2000 Calories.

Nutritional Intake (Calories)

2300	2000	2150	1950	2000	2150	1900	1900	2250	2050

Solution:

Step 1: Null hypothesis: The mean nutritional intake in the population, from which the sample is drawn, is 2000 Calories.

Alternative hypothesis: The mean nutritional intake in the population, from which the sample is drawn, is not 2000 Calories.

Step 2: Probability level for the hypothesis testing is 5%.

Step 3: Calculations:

Nutritional Intake (Calories)	d = X - A	d^2
2300	300	90000
2000	0	0
2150	150	22500
1950	-50	2500
2000	0	0
2150	150	22500
1900	-100	10000
1900	-100	10000
2250	250	62500
2050	50	2500
20650	650	222500

$n = 10$ $\quad \Sigma d^2 = 222500$ $\quad d_m = 650 / 10 = 65$

$M = 20650 / 10 = 2065$ $\quad \mu = 2000$

$S = v [\{\Sigma d^2 - (d_m)^{2*}n\} / (n - 1)]$

$S = v [\{222500 - (65)^{2*}10\} / 9 = 141.52$

$t = \{(M - \mu) * vn\} / S = \{(2065 - 2000) * v (10)\} / 141.52 = 1.452$

Step 4: Degree of freedom = 10 -1 = 9

Step 5: The table value of t at 5% probability level for 9 degree of freedom is = 2.232. The calculated value of t (1.452) is lower than the table value of t (2.232).

Thus the null hypothesis is accepted and the difference is not significant.

b) To test the difference between the means of the two independent samples: This test is used when the researcher is interested in examining whether the respective means of two independent samples differ significantly from each other. The following formula is used for its calculation:

$$t = [(M_1 - M_2) * v \{(n_1 * n_2) / (n_1 + n_2)\}] / S$$

When using actual mean:

$$S = v [\{\Sigma(X_1 - M_1)^2 + \Sigma(X_2 - M_2)^2\} / (n_1 + n_2 - 2)]$$

When using assumed mean

$$S = v [\{\Sigma d_1^2 + \Sigma d_2^2 - n_1 (M_1 - A_1)^2 - n_2 (M_2 - A_2)^2\} / (n_1 + n_2 - 2)]$$

Where $d_1 = X_1 - A_1$ and $d_2 = X_2 - A_2$ respectively

M_1 and M_2 are the respective means of the two samples

A_1 and A_2 are the assumed mean of the two samples

n_1 and n_2 are the sample sizes and S is the common standard deviation. We wuld take an example in Box 18.5 to find out the marital distance among the Santhals and Murias.

Box 18.5 Example to Examine whether Santhals and Murias Differ in Marital Distance

	Marital Distance (km)									
Santhals	10	12	15	17	18	17	19	22	22	12
Murias	22	19	21	23	18	21	23	20	19	21

Solution:

Step 1: Null hypothesis: Santhals and Murias do not differ in their marital distance.

Alternative hypothesis: Santhals and Murias differ in their marital distance.

Step 2: Probability level for the hypothesis testing is 5%.

Step 3: Calculations:

Santhals			Murias				
X_1	$d_1=X_1-A_1$	d_1^2	X_2	$d_2=X_2-A_2$	D_2^2	A1 =	A2=
10	-6	36	22	2	4	16	20
12	-4	16	19	-1	1	16	20
15	-1	1	21	1	1	16	20
17	1	1	23	3	9	16	20
18	2	4	18	-2	4	16	20
17	1	1	21	1	1	16	20
19	3	9	23	3	9	16	20
22	6	36	20	0	0	16	20
22	6	36	19	-1	1	16	20
12	-4	16	21	1	1	16	20
164	4	156	207	7	31		

$A_1 = 16$ $A_2 = 20$ $M_1 = 16.4$ $M_2 = 20.7$

$n_1 = 10$ $n_2 = 10$ $\Sigma d_1^2 = 156$ $\Sigma d_2^2 = 31$

$S = v [\{\Sigma d_1^2 + \Sigma d_2^2 - n_1 (M_1 - A_1)^2 - n_2 (M_2 - A_2)^2\} / (n_1 + n_2 - 2)]$

$S = v [\{156+ 31 - 10 (16.4 - 16)^2 - 10 (20.7 - 20)^2\} / (10 + 10 - 2)]$

$S = v [\{156+ 31 - 10 (16.4 - 16)^2 - 10 (20.7 - 20)^2\} / (10 + 10 - 2)]$

$S = v [180.5/ 18] = v10.028 = 3.167$

$t = [(M_1 - M_2) * v \{(n_1 * n_2) / (n_1 + n_2)\}] / S$

$t = \{(16.4 - 20.7) * v (100 / 20)\}/ 3.167 = (4.3 * 2.236) / 3.167 = 3.036$

Step 4: Degree of freedom = 10 + 10 - 2 = 18

Step 5: The table value of t at 5% probability level for 9 degree of freedom is = 2.101. The calculated value of t (3.036) is higher than the table value of t (2.101).

You can say that the null hypothesis is rejected and the difference in marital distance between Santhals and Murias is significant.

c) To test the difference between the means of the two dependent samples: This test is used when the researcher is interested in examining whether the mean of two dependent samples differ significantly from each other. The following formula is used for its calculation:

$$t = (d_m * v\, n) / S$$

$$S = v\, [\Sigma (d - d_m)^2 / (n - 1)] \text{ or}$$

$$S = v\, [(\Sigma d^2 - (d_m)^2 * n) / (n - 1)]$$

Where, $d = X_1 - X_2$;

d_m is the mean of the deviations;

n_1 and n_2 are the sample sizes; and

S is the common standard deviation. We would take an example in Box 18.6 to find out differences in observations of two researchers.

Box 18.6 Example: Two observers have taken the income of ten households. Examine whether their observations differ significantly?

	Household No.									
Observers	1	2	3	4	5	6	7	8	9	10
Observer 1	2400	1950	2200	1800	2050	2250	2000	1950	2300	2000
Observer 2	2300	2000	2150	1950	2000	2150	1900	1900	2250	2050

Solution:

Step 1: Null hypothesis: The difference in the observation by the two observers is not significant.

Alternative hypothesis: The difference in the observation by the two observers is significant.

Step 2: Probability level for the hypothesis testing is 5%.

Step 3: Calculations:

Household No.	Observer 1	Observer 2	$d = X_1 - X_2$	d^2
1	2400	2300	100	10000
2	1950	2000	-50	2500
3	2200	2150	50	2500
4	1800	1950	-150	22500
5	2050	2000	50	2500
6	2250	2150	100	10000
7	2000	1900	100	10000
8	1950	1900	50	2500
9	2300	2250	50	2500
10	2000	2050	-50	2500
			250	67500

n =10 $d_m = 250/10 = 25$

$S = v [\Sigma d^2 - (d_m)^2 * n / (n - 1)]$

$S = v [(67500 - (25)^2 * 10) / 9]$

S = 82.496

$t = (d_m * v n) / S$

t = (25* v 10) / 82.496 = 0.958

Step 4: Degree of freedom = 10 -1 = 9

Step 5: The table value of t at 5% probability level for 9 degree of freedom is = 2.232. The calculated value of t (0.958) is lower than the table value of t (2.232).

You would say that the null hypothesis is accepted and the difference between the observers is not significant.

d) To test the significance of Correlation coefficient: Whether a coefficient of correlation is significant or not may be tested using the following formula:

$$t = \frac{r * v (n - 2)}{v (1 - r^2)}$$

Where, r is the coefficient of correlation and n is the number of observations. Degree of freedom is n-2. In Box 18.7, we would take an example to test test the importancve of a correlation.

Box 18.7 Example: Using the Following Data to Test the Significance of the Correlation

r = 0.45, n = 102

Step 1: Null hypothesis: The coefficient of correlation is not significant.
Alternative hypothesis: The coefficient of correlation is significant.

Step 2: Probability level for the hypothesis testing is 5%.

Step 3: Calculations:

$t = (r * v (n - 2)) / v (1 - r^2)$

$t = (0.45 * v (100)) / v (1 - 0.45*0.45)$

$t = (0.45*10) / v (1 - 0.2025) = 4.5 / v0.7975 = 4.5 / 0.893 = 5.039$

Step 4: Degree of freedom = 102 -2 = 100

Step 5: The table value of t at 5% probability level for 100 degree of freedom is = 1.96. The calculated value of t (5.039) is higher than the table value of t (1.96).

Thus, you would find that the null hypothesis is rejected and the correlation is significant.

Reflection and Action 18.3

Of the four tests in section 18.4, select one test and carry it out with respect to your own research work. Write it out in detail in your research work report.

18.5 Conclusion

Unit 18 has provided you with a range of ways to draw inferences. There are a good number of examples given for you to try and prepare your own examples. The exercises of working with as many as possible examples would help you to master the skills of testing hypotheses and estimating unknown parameters of the population. You need to keep in mind that no matter what design you used to test a hypothesis, you would reach only approaximations in terms of probability. The testing of a hypothesis prepares for you the ground for generating further hypotheses and in this manner the scientific knowledge progress. Initial approaximations put on firm basis the original hypothesis and from this you can further deduce other hypotheses. If you are able to establish links between propositions you would have generated scientific knowledge.

Further Reading

Handel, J.D. 1978, *Statistics for Sociology*, Englewood Cliffs, N.J.

Watson, G. and McGawd 1980. *Statistical Inquiry: Elementary Statistics for the Political Science and Policy Sciences.* John Wiley: New York

Unit 19

Correlation and Regression

Contents

Learning Objectives

It is expected that after reading Unit 19 you would be able to

- ❖ Appreciate the relevance of the analysis of co-variation between two or more variables
- ❖ Describe different types of correlation
- ❖ Elaborate methods of calculating correlation of both ungrouped and grouped data
- ❖ Understand the method of regression analysis that helps in estimating the values of a variable from the knowledge of one or more variables.

19.1 Introduction

In the concluding Section of Unit 18, we mentioned the linkages between propositions. Let us now discuss the subject of correlation and regression.

Unit 19 is about correlation, that is an analysis of co-variation between two or more variables. You would notice that the statistical tool of correlation helps to measure and express the quantitative relationship between two variables. Unit 19 elaborates the ways of applying the tool. It shows the relevance of coefficient of correlation, coefficient of determination and regression analysis in the social sciences. Further, it explains regression analysis, which is the method of estimating the values of a variable from the knowledge of one or more variables. The unit tells you to use the statistical tool of correlation without fear or apprehension that its application is difficult and complex.

19.2 Correlation

Correlation[●] is an analysis of the co-variation between two or more variables. When the relationship between the two variables is quantitative, the statistical tool for measuring the relationship and expressing it in a brief formula is known as correlation. If a change in one variable results in a corresponding change in the other, the two variables are correlated. Let us look at types of correlation.

Types of correlation

Probing into the types of correlation, we contemplate two types : correlation:

A) Positive and Negative correlation;

B) Linear and Non-linear correlation

A) Positive and negative correlation

If the values of the two variables deviate in the same direction, i.e., if an increase in the value of one results on an average in a corresponding increase in the value of the other, or if decrease in the value of one variable results in a decrease in the value of the other, then correlation is said to be p*ositive or direct*. Some examples of a series of positive correlation are (i) height and weight (ii) land owned and household income. On the other hand, if the variables deviate in the opposite directions, i.e. if an increase (decrease) in the value of one variable, on an average, results in a decrease (increase) in the value of the other variable, then the correlation is n*egative or indirect*. Some examples of negative correlation are (i) physical assets and the level of poverty, (ii) muscle strength and age. Figure 19.1 shows the positive and negative types of correlation.

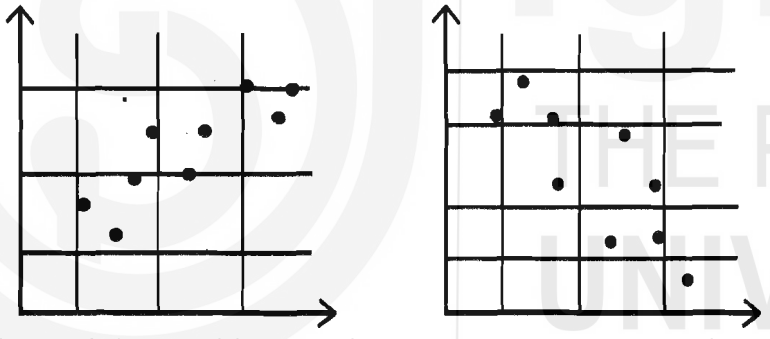


Figure 19.1 (a) Positive Correlation and (b) Negative Correlation

The values of correlation range from -1 to +1.When r = +1, it means there is perfect positive correlation between the variables. When r = -1, there is perfect negative correlation. When r = 0, it means there is no correlation between the two variables (see Figure 19.2).

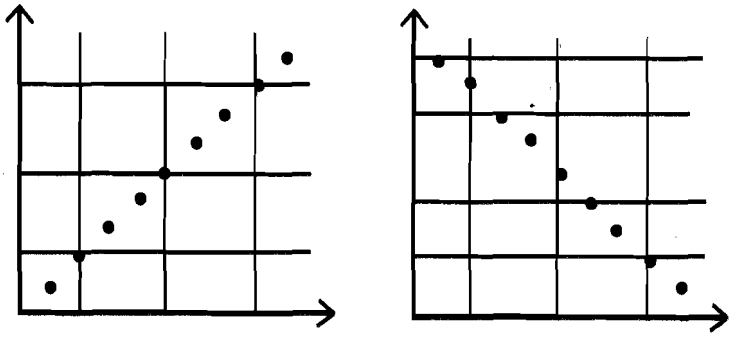


Figure 19.2 (a) Perfect Positive Correlation (r=+1) and (b) Perfect Negative correlation (r =-1)

B) Linear and non-linear correlation

The correlation between two variables is said to be linear if corresponding to a unit change in one variable, there is a constant change in the other variable over the entire range of the values. Consider the following data in Figure 19.3.

X	1	2	3	4	5	6
Y	3	5	7	9	11	13

Figure 19.3 Constant Change Figuring in the Entire Range of Values

In this case, the data in Figure 19.3 can be represented by the relation Y=1 + 2 X. In general, two variables are said to be linearly related if there exists a relationship of the form Y=a + b X.

On the other hand, the relationship between the two variables is said to be non-linear or curvilinear if corresponding to a unit change in one variable, the other variable does not change at a constant but a fluctuating rate. Example of a non-linear correlation is given by the following data set in Figure 19.4.

X	1	2	3	4	5	6
Y	5	8	14	15	18	22

Figure 19.4 Non-linear Correlation

In the example in Figure 19.4, there is fluctuating (not constant) change in the value of Y corresponding to a unit change in the value of X, and thus it represents a non-linear correlation.

You would like to know how to study correlation. Let us briefly discuss the methods of studying correlation. But before going on to metods of studying correlation, let us complete Reflection and Action19.1

Reflection and Action 19.1

Relating to your hypothesis, draw the figure of its positive and negative correlations. Next draw another figure of perfect positive and perfect negative correlations. In addition, draw two more figures of constant change reflected in the entire range of values and non-linear correlation. You may take help of Figures 18.1 to 18.4 in the text above for drawing your figures.

Methods of studying correlation

The various methods to determine whether there is a correlation between two variables are (i) Scatter diagram; (ii) Graphic method; (iii) Karl Pearson's coefficient of correlation; (iv) Rank method; (v) Concurrent deviation method; and (vi) Method of least squares. Of these, the first two are based on the knowledge of diagrams and graphs and the rest on

mathematical tools. Of the several mathematical tools used, the most popular is the Karl Pearson coefficient of correlation (r) and thus we will focus on this method. The procedure is different for calculating correlation from ungrouped and grouped data.

19.3 Method of Calculating Correlation of Ungrouped Data

There are various methods for the calculation of the coefficient© of correlation from ungrouped data.

i) Using actual mean
ii) Using assumed mean
iii) Direct method

The use of all these methods is illustrated with the help of the following example.

Example: Find out the correlation coefficient (Karl Pearson's) between the age at marriage of husbands and wives using the following data in Figure 19.5

Age at Marriage	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7	Case 8	*Case 9*	*Case 10*
Husbands	28	25	24	29	31	22	21	25	26	28
Wives	22	23	21	25	26	20	19	21	21	24

Figure 19.5 Correlation Coefficient between the Age at Marriage of Husbands and Wives

Method of calculating correlation coefficient using the actual mean

You would first learn the method of calculating correlation coefficient using the actual mean and then you would actually carry out the calculation itself.

The formula used for calculating r is:

$$r = \Sigma xy / N * ó_x * ó_Y$$

Where, $x = (X - M_x)$ in which M_x is the mean of series of X values;

$y = (Y - M_Y)$ in which M_y is the mean of series of Y values;

$ó_x$ = Standard deviation of series X

$ó_Y$ = Standard deviation of series Y

N = Number of pair of observations

This formula can also be expressed as:

$$r = \Sigma xy / \Sigma [Óx^2 * Óy^2]$$

The following steps elucidate the calculation of the coefficient of correlation.

I. Take deviations of X series from the mean of X and denote them by x;

II. Square these deviations and obtain the total, i.e. Óx²;

III. Take deviations of Y series from the mean of Y and denote them by y;

IV. Square these deviations and obtain the total, i.e. Óy²;

V. Multiply the deviations of x and y and obtain the total Ó xy; and

VI. Substitute the values of Óx², Óy², and Ó xy in the above formula.

Calculation of correlation coefficient using actual mean

After learning the method, let us now make the calculation as reflected in Figure 19.6

X	$x = X - M_x$	X^2	Y	$y = Y - M_y$	y^2	Xy
28	2.1	04.41	22	-0.2	00.04	-00.42
25	-0.9	00.81	23	0.8	00.64	-00.72
24	-1.9	03.61	21	-1.2	01.44	02.28
29	3.1	09.61	25	2.8	07.84	08.68
31	5.1	26.01	26	3.8	14.44	19.38
22	-3.9	15.21	20	-2.2	04.84	08.58
21	-4.9	24.01	19	-3.2	10.24	15.68
25	-0.9	00.81	21	-1.2	01.44	01.08
26	0.1	00.01	21	-1.2	01.44	-00.12
28	2.1	04.41	24	1.8	03.24	03.78
259	0	88.90	222	0	45.60	58.20

Figure 19.6 Calculation of Correlation Coefficient using Actual Mean

$r = \Sigma xy / ? [Óx^{2*} Óy^2]$

$M_x = 259 / 10 = 25.9$ $M_y = 222 / 10 = 22.2$ $(Óx^2) = 88.9$ $(Ó y^2) = 45.6$ Ó xy = 58.2

$r = 58.2 / v [88.9 * 45.6] = 0.914$

Method of calculating correlation coefficient using assumed mean

The only difference in this method as compared to the above method is that in the former, the deviations are taken from the actual mean, and in this case from the assumed mean (i.e. by looking at the series of X and Y, assume means for X and Y and proceeding in the same manner).

Calculation of correlation coefficient using assumed mean

You would now calculate as per Figure 19.7.

X	D_x=X - Ax	d_x^2	Y	d_Y = Y- Ay	d_Y^2	d_x*d_Y
28	3	9	22	0	0	0
25	0	0	23	1	1	0
24	-1	1	21	-1	1	1
29	4	16	25	3	9	12
31	6	36	26	4	16	24
22	-3	9	20	-2	4	6
21	-4	16	19	-3	9	12
25	0	0	21	-1	1	0
26	1	1	21	-1	1	-1
28	3	9	24	2	4	6
259	9	97	222	2	46	60

Figure 19.7 Calculation of correlation coefficient using assumed mean

$$r = \frac{N\Sigma d_x * d_Y - (? d_X * ? d_Y)}{v\{N \Sigma d_x^2 - (\Sigma d_x)^2\} * v\{N \Sigma d_y^2 - (\Sigma d_y)^2}$$

$$r = \frac{10*60 - (9* 2)}{v\{10*97 - (9)^2\} * v\{10*46 - (2)^2}$$

$$r = \frac{582}{636.697}$$

r = 0.914

Direct method of calculating correlation coefficient

The coefficient can also be calculated by taking actual X and Y values, without taking deviations either from the actual or assumed mean. The formula for its calculation is as follows.

$r = (N * \Sigma XY - \Sigma X * \Sigma Y) / v [N * \Sigma X^2 - (\Sigma X)^2] * v [N * \Sigma Y^2 - (\Sigma Y)^2]$

The direct method gives the same answer as one gets when deviations are taken from the assumed or actual means. The example demonstrates this point in Figure 19.8.

X	Y	X^2	Y^2	XY
28	22	784	484	616
25	23	625	529	575
24	21	576	441	504
29	25	841	625	725
31	26	961	676	806
22	20	484	400	440
21	19	441	361	399
25	21	625	441	25
26	21	676	441	546
28	24	784	576	672
259	222	6797	4974	5808

Figure 19.8 Calculation of correlation coefficient using direct method

Let us now complete Reflection and Action 19.2 and then learn in Section 19.4 the methods of calculating correlation of grouped data.

Reflection and Action 19.2

Select one of the following two calculations and carry it out in relation to your hypothesis. You need not worry about making mistakes in your calculations. At the moment the idea is to learn the procedure. This is not to be a part of your report.

i) Calculation of correlation coefficient using assumed mean

ii) Calculation of Correlation Coefficient using Direct Method

19.4 Method Of Calculating Correlation Of Grouped Data

With a large number of observations, the data is concealed into a two-way frequency distribution called correlation table. The class intervals of Y series are written as column headings and that of the X series are written as row headings. The frequency distribution for the two variables is written in the respective cells. The formula for calculating the coefficient of correlation is:

$$r = \frac{\Sigma f\,{}^{*}d_X{}^{*} d_Y - (\Sigma f_X{}^{*}d_X \,{}^{*} \Sigma f_Y{}^{*}d_Y) / N}{\sqrt{\{\Sigma f_X{}^{*}d_X^2 - (\Sigma f_X{}^{*}d_X)^2 / N\}^{*} \sqrt{\{\Sigma f_Y{}^{*} d_Y^2 - (\Sigma f_Y{}^{*}d_Y)^2 / N\}}}}$$

Steps:

i) Take the step deviations of variable X and denote these deviations by d_X

ii) Take the step deviations of variable Y and denote these deviations by d_Y

iii) Multiply $d_X \,{}^{*}d_Y$ and the respective frequencies for each cell and write the figure obtained in the right hand upper corner of the cell.

iv) Add together all values to obtain $\Sigma f\,{}^{*}d_X{}^{*} d_Y$

v) Multiply all the frequencies of the variable X by the deviations of X and obtain the total $\Sigma f_X{}^{*}d_X$

vi) Take the squares of the deviations of the variable X and multiply by respective frequencies to obtain $\Sigma f_X{}^{*}d_X^2$

vii) Multiply all the frequencies of the variable Y by the deviations of Y and obtain the total $\Sigma f_Y{}^{*}d_Y$

viii) Take the squares of the deviations of the variable Y and multiply by respective frequencies to obtain $\Sigma f_Y{}^{*}d_Y^2$

ix) Substitute the values for $\Sigma f_Y{}^{*}d_Y^2$, $\Sigma f_Y{}^{*}d_Y$, $\Sigma f_X{}^{*}d_X^2$, $\Sigma f_X{}^{*}d_X$, $\Sigma f\,{}^{*}d_X{}^{*}d_Y$ in the above formula to get the value of r.

Let us now take an example to calculate the Karl Pearson's coefficient of correlation using the data in Figure 19.9.

Expenditure on Luxury Items	(Income in Thousand Rs.)				
	20 - 25	25 - 30	30 - 35	35 - 40	40-45
0 - 4	28	12	05		
4 - 8	41	22	09	03	
8 - 12	09	33	28	14	16
12 -16		18	22	29	37
16-20			03	09	12

Figure 19.9 Coefficient Correlation regarding Expenditure on Luxury Items

We can calculate correlation coefficient in grouped data using direct method as seen in Figure 19.10 (See figure 19.10).

Expenditure on Luxury Items	Income in Thousand Rs.)								
	20 - 25	25 - 30	30 - 35	35 - 40	40-45	fy	DY	fy*dy	fy*dy*dy
0 - 4	28	12	5			45	-2	-90	180
4 - 8	41	22	9	3		75	-1	-75	75
8 - 12	9	33	28	14	16	100	0	0	0
12 -16		18	22	29	37	106	1	106	106
16 -20			3	9	12	24	2	48	96
Fx	78	85	67	55	65	350		-11	457
dx	-2	-1	0	1	2				
fx*dx	-156	-85	0	55	130	-56			
fx*dx*dx	312	85	0	55	260	712			

Figure 19.10 Calculation of Correlation Coefficient in Grouped Data

Now we can proceed to calculate fx*dx*dy using direct method as given in Figure 19.11.

Expenditure on Luxury Items	Income in Thousand Rs.)					fx*dx*dy
	20 - 25	25 - 30	30 - 35	35 - 40	40-45	
0 - 4	112	24	0	0	0	136
4 - 8	82	22	0	-3	0	101
8 - 12	0	0	0	0	0	0
12 -16	0	-18	0	29	74	85
16 - 20	0	0	0	18	48	66
fx*dx*dy	194	28	0	44	122	388

Figure 19.11 Calculation of Correlation Coefficient of Grouped Data

$N = 350 \qquad \Sigma f * d_X * d_Y = 388 \qquad \Sigma f_X * d_X = -56$

$\Sigma f_Y * d_Y = -11 \qquad \Sigma f_X * d_X^2 = 712 \qquad \Sigma f_Y * d_Y^2 = 457$

$$r = \frac{\Sigma f * d_X * d_Y - (\Sigma f_X * d_X * \Sigma f_Y * d_Y) / N}{}$$

$$v \{\Sigma f_x * d_x^2 - (\Sigma f_x * d_x)^2 / N\} * v \{\Sigma f_y * d_y^2 - (\Sigma f_y * d_y)^2 / N\}$$

$$r = \frac{388 - (-56 * -11) / 350}{v \{712 - (-56)^2 / 350\} * v \{457 - (-11)^2 / 350\}}$$

$$r = 386.24 / (26.515 * 21.369) = .682$$

Most of the variables show some kind of relationship. With the help of correlation one can measure the degree of relationship between two or more variables. Correlation, however, does not tell us anything about the cause and effect relationship. Even a high degree of relationship does not necessarily imply that a cause and effect relationship exists. Conversely, however the cause and effect relationship (or functional relationship) would always result in the expression of correlation.

We would now discuss regression analysis.

19.5 Regression

Regression© analysis is the method of estimating the values of a variable from the knowledge of one or more variables. The variable that the researcher tries to estimate is called dependent variable (denoted as Y), whereas the variable used for prediction is independent variable (denoted as X). In a regression equation, there may be one or more independent variables, but there is only one dependent variable. Depending on whether there are one or more independent variables, the regression equation is called simple or multiple. The term 'linear' is added if the relationship between the dependent and the independent variable is linear. Thus a simple linear regression equation is represented as

$$Y = a + b X$$

Where, Y is dependent variable

X is independent variable

'a' is regression constant

'b' is regression coefficient. It measures the change in Y corresponding to a change in X.

Similarly a multilinear regression equation is represented as

$$Y = a + b_1 X_1 + b_2 X_2 + \ldots b_n X_n$$

Where, Y is dependent variable

$X_1, X_2, \ldots X_n$, are independent variables

'a' is regression constant

'$b_1, b_2, \ldots b_n$' are respective regression coefficients.

Like the calculation of coefficient of the correlation, there are various methods of calculating regression equation:

1. From actual mean values of X and Y.

2. From assumed mean values of X and Y.

Calculation of regression equation using actual mean

Regression equation (of Y on X) can be calculated using the following formula:

$Y- M_Y = b_{yx} * (X- M_X)$ or

$Y- M_Y = r (ó_Y / ó_X) * (X- M_X)$

As, $b_{yx} = r (ó_Y / ó_X) = (?xy / ?x^2)$, the regression equation may be calculated using the following formula.

$Y- M_Y = (\Sigma xy / \Sigma x^2) * (X- M_X)$

Where, Y and X are dependent and independent variables respectively;

M_Y and M_X are means of Y and X variable respectively; and

$y = Y-M_Y$ and $x = X-M_X$

The following example illustrates the calculation of the regression equation.

Example: Calculate the regression equation using the following data, taking age at marriage of husbands as independent variable and that of wives as dependent variable (see Figure 19.11)

Age at Marriage	*Case 1*	*Case 2*	*Case 3*	*Case 4*	*Case 5*	*Case 6*	*Case 7*	*Case 8*	*Case 9*	*Case 10*
Husbands	28	25	24	29	31	22	21	25	26	28
Wives	22	23	21	25	26	20	19	21	21	24

Calculation of regression equation using actual mean (see Figure 19.12)

Age of Wives Y	$y = Y-M_Y$	y^2	Age of Husbands X	$x = X-M_X$	x^2	xy
22	-0.2	00.04	28	2.1	4.41	-0.42
23	0.8	00.64	25	-0.9	0.81	-0.72
21	-1.2	01.44	24	-1.9	3.61	02.28
25	2.8	07.84	29	3.1	9.61	08.68
26	3.8	14.44	31	5.1	26.01	19.38
20	-2.2	04.84	22	-3.9	15.21	8.58
19	-3.2	10.24	21	-4.9	24.01	15.68
21	-1.2	01.44	25	-0.9	0.81	01.08
21	-1.2	01.44	26	0.1	0.01	-0.12
24	1.8	03.24	28	2.1	4.41	03.78
222	0	45.60	259	0	88.9	58.20

Figure 19.12 Calculation of Regression Equation using Actual Mean

$M_Y = 222 / 10 = 22.2$ $M_X = 259 / 10 = 25.9$

$Y - M_Y = (\Sigma xy / \Sigma x^2) * (X - M_X)$

$Y - 22.2 = (58.2 / 88.9) * (X - 25.2)$

$Y - 22.2 = 0.655 * (X - 25.2)$

$Y - 22.2 = 0.655X - 16.96$

$Y = 5.24 + 0.655X$

Calculation of regression equation using assumed mean (see Figure 19.13)

Regression equation (of Y on X) can be calculated using the following formula, taking the assumed mean:

$Y - M_Y = b_{yx} * (X - M_X)$

Where, $b_{yx} = [\Sigma d_X * d_Y - (\Sigma d_X * \Sigma d_Y) / N] / [\Sigma d_x^2 - (\Sigma d_x)^2 / N]$

Y and X are dependent and independent variable respectively;

M_Y and M_X are mean of Y and X variables respectively

$d_Y = Y - AM_Y$ **and** $d_X = X - AM_X$

AM_Y and AM_X are the assumed mean of Y and X variable respectively; and

Calculation of regression equation using assumed mean

Age of Wives Y	$d_Y = Y - AM_Y$	dy^2	Age of Husbands X	$d_X = X - AM_X$ x	dx^2	dx * dy
22	0	0	28	3	9	0
23	1	1	25	0	0	0
21	-1	1	24	-1	1	1
25	3	9	29	4	16	12
26	4	16	31	6	36	24
20	-2	4	22	-3	9	6
19	-3	9	21	-4	16	12
21	-1	1	25	0	0	0
21	-1	1	26	1	1	-1
24	2	4	28	3	9	6
222	2	46	259	9	97	60

Figure 19.13 Calculation of Regression Equation using Assumed Mean

$M_Y = 222 / 10 = 22.2$ $M_X = 259 / 10 = 25.9$

$b_{yx} = [\Sigma d_X * d_Y - (\Sigma d_X * \Sigma d_Y) / N] / [\Sigma d_x^2 - (\Sigma d_x)^2 / N]$

$b_{yx} = [60 - (9*2) / 10] / [97 - 9*9/10]$

$b_{yx} = 58.2 / 88.9 = 0.655$

$Y - M_Y = b_{yx} * (X - M_X)$

$Y - 22.2 = 0.655 * (X - 25.2)$

$Y - 22.2 = 0.655X - 16.96$

$Y = 5.24 + 0.655X$

Standard error of estimate: Perfect prediction, using a regression equation is not possible (except when correlation value is -1 or + 1). Thus the researcher is interested in finding the accuracy of estimation of a regression equation. Standard error of estimate measures the error involved in using a regression equation as a basis of estimation. It can be calculated using the following equation:

$SEE_{y..x} = v \{\Sigma(Y - Y_c)^2 / N - 2\}$

Where, $SEE_{y..x}$ is Standard error of estimate

Y is dependent variable

Y_c is predicted value of Y

N is the number of observations

It can also be calculated from the following formula

$SEE_{y..x} = v \{(\Sigma Y^2 - a?Y - b\Sigma XY) / N- 2\}$

Where, $SEE_{y..x}$ is Standard error of estimate

Y is dependent variable

X is independent variable

'a' is regression constant

'b' is regression coefficient.

N is the number of observations

Coefficient of determination: Coefficient of determination (r^2) is the square of correlation coefficient (r) and is often used in interpreting the value of the coefficient of correlation. If the value of r were 0.8 then the coefficient of determination or r^2 would be 0.64. This would mean that 64% of variance of one variable (dependent) is explained in terms of the other variable (independent).

Reflection and Action 19.3

I tried to understand how to make the calculation of regression equation using assumed mean. I could not succeed. May be you can explain it to me with an example. Write out on a separate sheet of paper your explanation with one or two examples. May be I will then follow it. You will need to send it to the co-ordinator of MSO 002.

19.6 Conclusion

Unit 19 is the last unit of Block 5 on Quantitative Methods. All five units of this block have emphasised that quantitative methods should be used in social research when they are necessary and relevant and can provide superior results. Sometimes you can use them in combination with the qualitative methods. You need not avoid the quantitative methods because

of lack of information or apprehension that it is difficult to understand them. The five units of block 5 have provided you appropriate examples wherever possible and necessary to help you understand the tools that are very useful in your research project assignment.

Further Reading

Burns, Robert B. 2000. *Introduction to Research Methods*. Sage Publications: London

Cohen, Louis and Michael Holliday 1982. *Statistics for Social Research*. Harper and Row: London