UNIT 14 INDEPENDENCE OF ATTRIBUTES

Structure

14.1 Introduction
Objectives
14.2 Consistency of Data
14.3 Conditions of Consistency of the Data
14.4 Independence of Attributes
Criterion of Independence
14.5 Summary
14.6 Solutions / Answers
Glossary

14.1 INTRODUCTION

In Unit 13, you have seen that statistics which deals with the measurement of variables that can be broadly classified as quantitative and qualitative. Quantitative variables are those whose magnitude can be measured numerically. For example, income, height, weight of a group of individuals or number of laborers getting particular amount of wage, etc. Qualitative variables are those whose magnitude cannot be directly measured. An investigator can only study the presence or absence of particular quality in a group. Examples of such variables are sickness, insanity, extravagance, etc. We have also discussed the statistical methodology used for the analysis of quantitative data. You must have also noted that these qualitative variables are called attributes and theory of attributes deals with the measurement of data whose magnitude cannot be directly measured numerically. Though, the qualitative data can be quantified but for the sake of clear understanding and convenience, the statistical methodologies for the analysis of qualitative data have been separately developed. By reading Unit 13, you must now be familiar with the notations, terminology and concepts that are pre–requisites for proceeding any further.

In the present unit we will discuss consistency of the data, the conditions for consistency and the independence of attributes. Section 14.2 deals with the idea of consistency of data. A data is said to be consistent if no class frequency turns out to be negative. Section 14.3 discusses the conditions for consistency of the data. The conditions will be obtained in terms of ultimate class frequencies (already discussed in Unit 13). Section 14.4 illustrates the independence of the attributes i.e. we will study whether or not there is relationship of any kind between two attributes say A and B.

In consistent data no class frequency is negative

Objectives

After reading this unit, you should be able to

- check whether the data is consistent or not;
- describe the conditions for consistency of the data;
- explain the independence of the attributes; and

- test if there exists any relationship of any kind between two attributes or they are independent.

14.2 CONSISTENCY OF DATA

It is a well known fact that no frequency can be negative. If the frequencies of various classes are counted and any class frequency obtained comes out to be negative, then the data is said to be inconsistent. Such inconsistency arises due to wrong counting, or inaccurate addition or subtraction or sometimes due to error in printing. In order to test whether the data is consistent, all the class frequencies are calculated and if none of them is found to be negative, the data is consistent. It should be noted that if the data is consistent it does not mean that the counting is correct or calculations are accurate. But if the data is inconsistent, it means that there is either mistake or misprint in figures.

Data is inconsistent if frequency of attribute or combination is greater than total frequency N.

In order to test the consistency of data, obtain the ultimate class frequencies. If any of them is negative, the data is inconsistent. It would also be seen that no higher order class could have a greater frequency than the lower order class frequency. If any frequency of an attribute or combination of attributes is greater than the total frequency N (frequency of zero order), the data is inconsistent. The easy way to check whether the ultimate class frequencies are negative or not (i.e. checking the data for consistency), is to enter the class frequencies in the chart given in the Section 13.6 of the Unit 13. This will present an overall picture of all the ultimate class frequencies.

It is also possible to lay down conditions for consistency of data.The following section deals with the rules for testing the consistency of the data.

14.3 CONDITIONS FOR CONSISTENCY OF THE DATA

Condition 1: If there is only one attribute A

(i) $(A) \geq 0$

(ii) $(\alpha) \geq 0 \Rightarrow (A) \leq N$ since $N = (A) + (\alpha)$

Condition 2: If there are two attributes A and B then

(i) $(AB) \geq 0$ otherwise (AB) would be negative

(ii) $(AB) \geq (A) + (B) - N$ otherwise $(\alpha\beta)$ would be negative

Proof: We have

$$(\alpha\beta) = (\alpha) - (\alpha B)$$
$$= N - (A) - [(B) - (AB)]$$
$$(\alpha\beta) = N - (A) - (B) + (AB)$$
$$\therefore (AB) = (A) + (B) - N + (\alpha\beta)$$

Now if (AB) is less than (A) + (B) – N, then $(\alpha\beta)$ would be negative.

(iii) $(AB) \leq (A)$ otherwise $(A\beta)$ would be negative since

$$(A) = (AB) + (A\beta)$$

(iv) $(AB) \leq (B)$ otherwise (αB) would be negative since

$(B) = (AB) + (\alpha B)$

Condition 3: If there are three attributes A, B and C then

(i) $(ABC) \geq 0$ otherwise (ABC) would be negative.

(ii) $(ABC) \geq (AB) + (AC) - (A)$ otherwise $(AB\gamma)$ would be negative.

(iii) $(ABC) \geq (AB) + (BC) - (B)$ otherwise (αBC) would be negative.

(iv) $(ABC) \geq (AC) + (BC) - (C)$ otherwise $(A\beta C)$ would be negative.

(v) $(ABC) \leq (AB)$

(vi) $(ABC) \leq (AC)$

(vii) $(ABC) \leq (BC)$

(viii) $(ABC) \leq (AB) + (AC) + (BC) - (A) - (B) - (C) + N$ otherwise $(\alpha\beta\gamma)$ would be negative.

Proof: Relation (ii) of (3) is obtained in the following way

Since, $(AB\gamma) \leq (A\gamma)$

i.e. $(AB) - (ABC) \leq (A) - (AC)$

i.e. $(ABC) \geq (AB) + (AC) - (A)$

Similarly, other relations can be computed. Now (i) and (viii) give

$$(AB) + (AC) + (BC) \geq (A) + (B) + (C) - N \quad \ldots (1)$$

(ii) and (vii) give

$$(AB) + (AC) - (BC) \leq (A) \quad \ldots (2)$$

(iii) and (vi) give

$$(AB) - (AC) + (BC) \leq (B) \quad \ldots (3)$$

(iv) and (v) give

$$(AB) + (AC) + (BC) \leq (C) \quad \ldots (4)$$

Expressions given by equations (1), (2), (3), (4) are the conditions of consistency which are of course obtained from (i) - (viii) conditions of non negativity of class frequencies.

Example 1: Examine the consistency of the following data N = 1000, (A)= 800, (B) = 400, (AB) = 80, the symbols having their usual meaning.

Solution: We have

$$(\alpha\beta) = N - (A) - (B) + (AB)$$

$$= 1000 - 800 - 400 + 80 = -120$$

Since $(\alpha\beta) < 0$, the data is inconsistent.

Example 2: In a locality having a population of 1000 persons, 750 were males out of whom 530 were married. Among females the number of married ones were 350. Check the consistency of the data.

Solution: Let A represent Males, α represent Females, B represent Married and β represent Unmarried

Given N = 1000, (A) = 750, (AB) = 530 and $(\alpha\beta) = 350$

$(\alpha) = N - (A) = 1000 - 750 = 250$

$(\alpha) = (\alpha B) + (\alpha\beta)$

$(\alpha\beta) = (\alpha) - (\alpha B) = 250 - 350 = -100$

Since $(\alpha\beta) < 0$, the data are inconsistent

Example 3: If all A's are B's and all B's are C's show that all A's are C's.

Solution: Given (AB) = (A), (BC) = (B)

We have to prove (AC) = (A)

We have the relation

$$(AB) + (BC) - (AC) \leq (B)$$

$$(A) + (B) - (AC) \leq (B)$$

$$\Rightarrow (A) \leq (AC)$$

But we know that (A) cannot be less than (AC), hence (A) = (AC)

Example 4: Among the adult population of a certain town 50% of the population is male, 60% are wage earners and 50% are 45 years of age or over. 10% of the males are not wage earners and 40% of the males are under 45. Can we infer anything about what percentage of the population of 45 or over are wage earners?

Solution: Let A, B, C denote the attributes male, wage earners and 45 years old respectively.

Then, N = 100, (A) = 50, (B) = 60, (C) = 50

$$(A\beta) = \frac{10}{100} \times 50 = 5,$$

$$(A\gamma) = \frac{40}{100} \times 50 = 20$$

We are to find out the limits of (BC)

$$(AB) = (A) - (A\beta) = 45,$$

$$(AC) = (A) - (A\gamma) = 30$$

Applying the conditions of consistency

(i) $(AB) + (AC) + (BC) \geq (A) + (B) + (C) - N$

$$\Rightarrow (BC) \geq -15$$

(ii) $(AB) + (AC) - (BC) \leq (A)$

$$\Rightarrow (BC) \geq 25$$

(iii) $(AB) - (AC) + (BC) \leq (B)$

$$\Rightarrow (BC) \leq 45$$

(iv) $(ABC) - (AB) + (AC) + (BC) \leq (C)$

$$\Rightarrow (BC) \leq 65$$

(ii) to (iv) $\Rightarrow 25 \leq (BC) \leq 45$

Hence the percentage of wage earning population of 45 years or over must be between 25 and 45.

14.4 INDEPENDENCE OF ATTRIBUTES

Two attributes A and B are said to be independent if there does not exist any kind of relationship between them. Thus, if A and B are independent we may expect (i) the same proportion of A's in B's as in β's and (ii) same proportion of B's in A's as in α's or we can say two attributes A and B are independent if A is equally popular in B's and in β's and B is equally popular in A's and in α's. For example, intelligence and honesty are independent the proportion of intelligent persons among honest and dishonest person must be equal. If proportion of intelligent persons among honest persons is more than the proportion of intelligent persons among dishonest persons, then obviously intelligence and honesty are not independent. There exists an association between them.

14.4.1 Criterion of Independence

If two attributes are independent then (i) in Section 14.4 gives

$$\frac{(AB)}{(B)} = \frac{(A\beta)}{(\beta)} \qquad \ldots (5)$$

$$\Rightarrow 1 - \frac{(AB)}{(B)} = 1 - \frac{(A\beta)}{(\beta)}$$

$$\frac{(\alpha B)}{(B)} = \frac{(\alpha\beta)}{(\beta)} \qquad \ldots (6)$$

Similarly, condition (ii) in Section 14.4 gives

$$\frac{(AB)}{(A)} = \frac{(\alpha B)}{(\alpha)} \qquad \ldots (7)$$

$$\Rightarrow 1 - \frac{(AB)}{(A)} = 1 - \frac{(\alpha B)}{(\alpha)}$$

$$\therefore \frac{(A\beta)}{(A)} = \frac{(\alpha\beta)}{(\alpha)} \qquad \ldots (8)$$

Infact, equation (5) $\Leftrightarrow$ equation (7) i.e. equation (5) implies equation (7) and equation (7) implies equation (5).

Now, we know that for independence

$$\frac{(AB)}{(B)} = \frac{(A\beta)}{(\beta)} = \frac{(AB)+(A\beta)}{(B)+(\beta)} = \frac{(A)}{N}$$

Since (AB) + (Aβ) = (A) and (B) + (β) = N

$$\therefore (AB) = (A).\frac{(B)}{N} \qquad \ldots (9)$$

Relation (9) and the expressions, which can be derived like this, give the condition to test the independence of two attributes A and B. The relation (9) can also be written as

$$\frac{(AB)}{N} = \frac{(A)}{N} \cdot \frac{(B)}{N}$$

Thus, an important rule for judging the independence between two attributes A and B can be formulated in terms of proportion. We may say that for independence, the proportion of AB's in the population should be equal to the product of the proportions of A's and B's in the population. The criteria of the independence between two attributes would be more comprehensible and easily understood with the help of following table. In the Table 1, the class frequencies are displayed in the relevant cells.

Table 1

Attributes	**A**	**α**	**Total**
B	(AB)	(αB)	(B)
β	(Aβ)	(αβ)	β
Total	(A)	(α)	(N)

Observing the above table,we may obtain the condition of independence as

$$(AB) = \frac{(A)(B)}{N}$$

or

$$\frac{(AB)}{N} = \frac{(A)}{N}\frac{(B)}{N}$$

Now, let us solve the following exercise:

E1) Given the following class frequencies, do you find any inconsistency in the data?

(A) = 300; (B) = 150; (αβ) = 110; N = 500.

E2) In a survey of 1000 children, 811 liked pizza; 752 liked chowmein and 418 liked burger; 570 liked pizza and chowmein; 356 liked pizza and burger; 348 liked chowmein and burger; 297 liked all the three. Test the consistency of the data.

E3) In a competitive examination 200 graduates appeared. Following facts were noted.

No. of boys =139

No. of Science graduate girls who failed to qualify for interview = 25

No. of Arts graduate girls who qualified for interview = 30

No. of Arts graduate girls who failed to qualify for interview = 18.

Test the consistency of the data.

E4) If report gives the following frequencies as actually observed, show that there is misprint or mistake of some sort.
N=1000; (A) = 525; (B) = 485; (C) = 427; (AB) = 189; (AC) = 140; (BC) = 85.

E5) A study was made about the studying habits of the students of certain university and the following facts were observed. Of the student surveyed, 75% were from well to do families, 55% were boys and 60% were irregular in their studies out of irregular ones 50% were boys and 2/3 were from well to do families. The percentage of irregular boys from well to do families was 8. Is there any consistency in the data?

Before ending this unit let us go over its main points.

14.5 SUMMARY

In this unit, we have discussed:

1. The data is consistent if none of the class frequency is negative. Consistency of the data does not imply that counting of the frequencies or calculations are correct. But the inconsistency in the data means that there is somewhere error or misprint in figures;
2. There are certain conditions laid down to check the consistency of data. These must be applied at the very outset of analysis to get correct and measurable results from data; and
3. Two attributes A and B are independent if

$$(AB) = \frac{(A).(B)}{N}$$

$$\text{or } \frac{(AB)}{N} = \frac{(A)}{N}.\frac{(B)}{N}$$

14.6 SOLUTIONS / ANSWERS

E1) First find out (AB)

$\because (\beta) = (A\beta) + (\alpha\beta)$

$350 = (A\beta) + 110$

$\therefore (A\beta) = 350 - 110 = 240$

Now, $(A) = (AB) + (A\beta)$

$300 = (AB) + 240$

$\therefore (AB) = 60$

For two attributes A and B conditions for consistency are

(i) $(AB) \geq 0$

$60 > 0$

(ii) $(AB) \leq (A)$

$60 < 300$

(iii) $(AB) \leq (B)$

$60 < 150$

(iv) $(AB) \geq (A) + (B) - N$

$60 \geq 300 + 150 - 500 = -50$

Since, all the conditions are satisfied

$\therefore$ the data are consistent

E2) Let A, B, C represent liking of pizza, chowmein and burger respectively.

The given data are

$N = 1000$; $(A) = 811$; $(B) = 752$; $(C) = 418$;

$(AB) = 570$; $(AC) = 356$; $(BC) = 348$; $(ABC) = 297$

Appling the conditions for testing consistency of three attributes we find that the condition (viii) is not satisfied i.e.

$$(ABC) \leq (AB) + (AC) + (BC) - (A) - (B) - (C) + N$$

Otherwise, $(\alpha\beta\gamma)$ would be negative.

$$297 \leq 570 + 356 + 348 - 811 - 752 - 418 + 1000$$

$$= 2274 - 1981 = 293$$

But $(ABC) > 293$

Thus, the data are inconsistent as $(\alpha\beta\gamma)$ would be negative.

E3) Let A represents boys

α represents girls

B represents science graduates

β represents arts graduates

C those who qualified for interviews

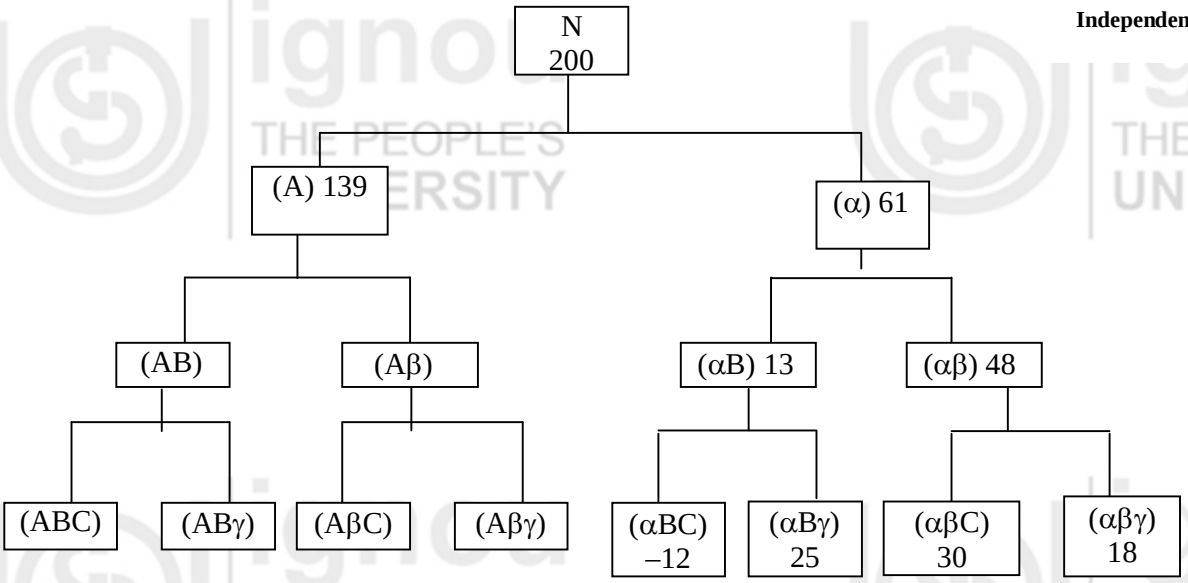
γ those who failed to qualify for interview

Hence, the given data are

$N = 200$; $(A) = 139$; $(\alpha B\gamma) = 25$; $(\alpha\beta\gamma) = 18$; $(\alpha\beta C) = 30$

In order to check the consistency of the data we have to find whether any ultimate class frequency is negative or not.

The easiest way is to enter the class frequencies in the chart given in the Section 13.6 of Unit 13, i.e.



∵ (αBC) is negative,

∴ The data are inconsistent.

E4) The condition of consistency when positive class frequencies are given (see equation (1) of Section 14.3)

$$(AB) + (AC) + (BC) \geq (A) + (B) + (C) - N \quad \text{or}$$

$$189 + 140 + 85 \geq 525 + 485 + 427 - 1000$$

$$414 \geq 437$$

which is not true.

Therefore, there is misprint or mistake of any sort in the report.

E5) Let A represent well to do families

B represents boys and C represents irregulars. The data given then are

$$N = 100;\ (A) = 75;\ (B) = 55;\ (C) = 60$$

$$(BC) = \frac{60 \times 50}{100} = 30\ ;\ (AC) = \frac{60 \times 2}{3} = 40\ ;\ (ABC) = 8$$

∵ (AB) is not given

∴ applying the (iv) condition of consistency of three attributes

$$(ABC) \geq (AC) + (BC) - (C)$$

$$8 \geq 40 + 30 - 60 = 10$$

which is not true.

Hence, data is inconsistent.

GLOSSARY

Consistency : Degree of firmness, reliably unchanging in deed, compatible.

Independence : No relationship of any kind.