Academic Journal of Statistics and Mathematics (AJSM)
Vol.6, No.10; October - 2020;
ISSN (5730 – 7151);
p –ISSN 4052 – 392X
Impact factor: 3.9

BIAS RESULTING FROM THE STUDY OF SELECTED SAMPLES

[1]Matthew Chukwuma Michael and [2]Oyeka I. C. A.

[1]Department of Mathematics and Statistics, Delta State Polytechnic, Ogwashi-Uku, Delta State, Nigeria. Email:
[2]Department of Statistics, Faculty of Physical Sciences, Nnamdi Azikiwe University, Awka, Anambra State.

Abstract: This study developed and presented the method for estimating the biases resulting from the study of selected samples in a population. The algebra underlying the method is a simple manipulation of conditional probabilities in a fourfold table. The entire goal of the study is divided into two namely; to develop the method for determining the proportion, out of those who possess the condition and are hospitalized, who possess the antecedent factor and; to develop the method for determining the proportion, out of those who possess the condition and are hospitalized, who do not possess the antecedent factor. Specifically the method was illustrated with a situation where not all subjects are equally like to end up in the study sample and bias results when the association found in the selected samples are presumed to apply in the population at large. It was found that the two probabilities are not equal if the rates of hospitalization are not, even though the corresponding probabilities of persons having the antecedent factor given that they have the condition and having the antecedent factor given that they do not have the condition in the community are equal.

Keywords: Antecedent factors; biases; conditional probabilities; specific rates; association; fourfold tables.

INTRODUCTION

The first clues to the association between diseases and antecedent conditions frequently come from the study of such selected samples as hospitalized patients and autopsy cases. Because not all subjects are equally likely to end up in the study samples, bias may result when the associations found in the selected samples are presumed to apply in the population at large. A classic example of this kind of bias occurs in a study by (Pearl, 1929). A large number of autopsy cases were cross-classified by the presence or absence of cancer and by the presence or absence of tuberculosis. A negative association between these two diseases was found, that is, tuberculosis was less frequent in autopsy cases with cancer than in cases without cancer. Pearl (1929) inferred that the same negative association should apply to live patients, and in fact acted on the basis of this inference by conceiving a study to treat terminal cancer patients with tuberculin (the protein of the tubercle bacillus) in the anticipation that the cancer would be arrested. What Pearl (1929) ignored is that, unless all deaths are equally likely to be autopsied, it is improper to extrapolate to live patients an association for live patients but, due to the differential selection of patients for autopsy, a strong association for autopsied cases.

The same kind of bias is possible whenever the chances of a subject's being included in the study sample vary. This has been pointed out by (Berkson, 1946, pp. 47 - 53; Berkson, The Statistical Study of Association between Smoking and Lung Cancer., 1955 July 27; Mainland, The Risk of Fallacious Conclusions from Autopsy Data of the Incidence of Diseases with Apllication to Heart Disease, 1953; Mainland, Elementary Medical Statistics, 1963; White, 1953; Mantel, 1959). The bias is illustrated using hypothetical data.

Academic Journal of Statistics and Mathematics (AJSM)
Vol.6, No.10; October - 2020;
ISSN (5730 – 7151);
p –ISSN 4052 – 392X
Impact factor: 3.9

Suppose that a research worker in a general hospital reviewing data for a number of patients, comes up with the fourfold table shown as Table 1.3

Table 1: Prior Living Arrangement by Diagnosis

	Prior Living Arrangement			
Diagnosis	Alone	With Family	Total	**Proportion Alone**
Neurotic	48 (n_{11})	52 (n_{12})	100 $(n_{1.})$	$0.48 = p_1$
Non-Neurotic	108 (n_{21})	696 (n_{22})	800 $(n_{2.})$	$0.13 = p_2$
Total	152 $(n_{.1})$	748 $(n_{.2})$	900 $(n_{..} = n)$	(p)

There is clearly an association between whether a patient is or is not neurotic and whether he did not live alone: the proportion of neurotics who had lived alone, $p_1 = \frac{n_{11}}{n_{1.}} = \frac{48}{100} = 0.48$, is nearly four times the proportion of non-neurotics who lived alone, $p_2 = \frac{n_{21}}{n_{2.}} = \frac{108}{800} = 0.125 = 0.135$. Would it however, be correct to conclude that type of living arrangement and neuroticism were associated in the community? Not necessarily. These two characteristics – Neuroticism and Living Arrangement – may be independent in the community and yet end up as associated in the hospital. This phenomenon is always possible when admission rates for people with different combinations of factors vary and can really be ruled out only when the disease in question almost always require care (example; leukaemia and other cancers).

The algebra underlying the phenomena is as follows. Let B denote the event that a person has the disease (in this case, a neurosis) and $\bar{B}$ the event that a person does not have the disease. Let $P(B)$ denote the proportion of all people in the community who have the disease and let $P(\bar{B}) = 1 - P(B)$ denote the proportion of all people who are free of the disease. Let A denote the event that a person lives alone and, $\bar{A}$ the event that a person lives with his family. Let $P(A)$ and $P(\bar{A}) = 1 - P(A)$ represent the corresponding proportions of neurotics who live alone and neurotics who live with their families. Let $P(AB)$ denote the proportion of all people in the community who both have the disease and lived alone, and assume that these two characteristics are independent in the community. Thus,

$$P(AB) = P(A)P(B) \quad 1$$

Let H denote the event that a person from the community is hospitalized for some reason or the other. Define $P(H/AB) =$ the proportion, out of all the people who have the disease and who live alone, who are hospitalized; $P(H/\bar{A}B) =$the proportion, out of all people who have the disease and who live with their families, who are hospitalized; and define $P(H/A\bar{B})$ and $P(H/\bar{A}\bar{B})$ similarly. Our problem is to evaluate, in terms of these probabilities,

$$P_1 = P(A/BH) \quad 2$$

that is, the proportion, out of all people who have the disease and are hospitalized, who live alone; and

$$P_2 = P(A/\bar{B}H) \quad 3$$

that is, the proportion, out of all people who do not have the disease and are hospital, who live alone.

We make use of the following version of the definition of a conditional probability:

$$P_1 = P(A/BH) = \frac{P(AB/H)}{P(B/H)} \quad 4$$

In Equation 4 the second condition, H (for hospitalization), remains a condition qualifying all probabilities.

The numerator of Equation 4 may be evaluated as

Academic Journal of Statistics and Mathematics (AJSM)
Vol.6, No.10; October - 2020;
ISSN (5730 – 7151);
p –ISSN 4052 – 392X
Impact factor: 3.9

$$P(AB/H) = \frac{P(H/AB).P(AB)}{P(H)}$$
$$= \frac{P(H/AB)P(A)P(B)}{P(H)} \quad 5$$

Because of the assumed independence of A and B. The denominator of Equation 4 is

$$P(B/H) = \frac{P(H/B).P(B)}{P(H)} \quad 6$$

To find $P(H/B)$, we make use of the fact that an overall rate is a weighted average of specific rates. This time, however, we apply the additional fact that a conditional probability, in this case B, remains one in all subsequent probabilities. Thus,

$$P(H/B) = P(H/AB)P(A/B) + P(H/\bar{A}B)P(\bar{A}/B)$$
$$= P(H/AB)P(A) + P(H/\bar{A}B)P(\bar{A})$$

because the assumed independence of A and B implies that $P(A/B) = P(A)$ and $P(\bar{A}/B) = P(\bar{A})$. Therefore Equation 6 becomes

$$P(B/H)$$
$$= \frac{P(B)\big(P(H/AB)P(A) + P(H/\bar{A}B)P(\bar{A})\big)}{P(H)} \quad 7$$

Substitution of Equation 5 for the numerator of Equation 4 and of the denominator, yields

$$P_1 = \frac{P(H/A\bar{B})P(A)}{P(H/AB)P(A) + P(H/\bar{A}B)P(\bar{A})} \quad 8$$

Similarly,

$$P_2 = \frac{P(H/A\bar{B})P(A)}{P(H/A\bar{B})P(A) + P(H/\bar{A}\bar{B})P(\bar{A})} \quad 9$$

These two probabilities of the rates of hospitalization are not equal, even though the corresponding probabilities in the community, $P(A/B)$ and $P(A/\bar{B})$ are equal.

As an example, suppose that 60% ($P_{.1}$) of all people live alone, so that $P(A) = P_{.1} = 0.60$ and suppose that the various hospitalization rates are $P(H/AB) = 0.25, P(H/\bar{A}B) = 0.40, P(H/A\bar{B}) = 0.01$ and $P(H/\bar{A}\bar{B}) = 0.10$. These rates are such that neurotics, whether they live with their families or they live alone have higher hospitalization rates than non-neurotics and that people who live with their families, both neurotics and non-neurotics, have higher hospitalization rates than people who live alone.

Substituting the values into Equations 8 and 9, we find that

$$P_1 = \frac{0.25 \times 0.60}{0.25 \times 0.60 + 0.40 \times 0.40} = \frac{0.15}{0.31} = 0.48$$

and that

$$P_2 = \frac{0.01 \times 0.60}{0.01 \times 0.60 + 0.10 \times 0.40} = \frac{0.006}{0.046} = 0.13$$

The results obtained here are consistent with those on Table 1.

3. CONCLUSION

The moral of this exercise is clear. Unless something is known about differential hospitalization rates on differential autopsy rate, a good amount of scepticism should be applied to any generalization from associations for people at large. This caveat obviously applies also to associations obtained from reports by volunteers

Conflict of Interest: On behalf of all the authors, I state that we have no conflict of interest.

References

Berkson, J. (1946). Limitations of the Application of Fourfold Table Analysis to Hospital Data. *Biometrics Bulletin, 2*(3), 47 - 53.

Berkson, J. (1955 July 27). The Statistical Study of Association between Smoking and Lung Cancer. *Proc. Staff Meet Clinic, 15*, pp. 319 - 348.

Berkson, J. (1955, July 27). The Statistical Study of Association between Smoking and Lung canceer. *Proc. Staff Meet Mayo Clinic, 30*(15), 319 -348.

Mainland, D. (1953). The Risk of Fallacious Conclusions from Autopsy Data of the Incidence of Diseases with Apllication to Heart Disease. *Amer. Heart J., 45*, 644 - 654.

Mainland, D. (1963). *Elementary Medical Statistics* (Second ed.). Philadelphia: W. W. Saunders.

Mantel, N. a. (1959). Statistical Aspect of the Analysis of Data from Retrospective Studies of Diseases. *J. Natl. Cancer Inst., 22*, 719 - 748.

Pearl, R. (1929). Cancer and Tuberculosis. *Amer. J. Hyg. (Now Amer. J. Epidemiol.), 9*, 97 - 159.

White, C. (1953). Sampling in Medical Research. *Brit. Med. J., 2*, 1284 - 1288.