DOI:10.5007/2175-7984.2011v10n19p57

Política & Sociedade - Florianópolis - Volume 10 - Nº 19 - outubro de 2011

Human ecology, statistical analysis, and the logic of valid correlations

*Fernando Dias de Avila-Pires**

I believe the universe is a great symphony of numerical correspondences, I believe that numbers and their symbolisms provide a path for special knowledge.
Umberto Eco (1990, p. 242)

Abstract

Human health ecology is an expanding field of interdisciplinary research, combining knowledge drawn from distinct areas. Human ecology is essentially interdisciplinary. The search for the origin and distribution of diseases demands a combination of theories and methods from both natural and social sciences. There is a limit to the replication and repetition of observations and experiments in scientific research: hence the importance of statistics. As a tool, statistical analyses are usually based upon selected samples and samples must be representative of the whole. Irrelevant, erroneous, false, and misleading, information based on flawed interpretation of statistics has always plagued scientific publications. In some cases, not only professional ethics has been breached, but risks to well-being and health has been fostered. In this article I will not deal with fraud, but with *bona fide* errors of judgment and the use of erroneous statistical concepts. My objective in this article is to caution against a common practice adopted by researchers while using qualitative as well as quantitative methods of analysis to establish demographic categories.

Keywords: human ecology, correlations, statistical analysis, categories, confounding variables.

* Pesquisador Titular (aposentado) do Departamento de Medicina Tropical, Instituto Oswaldo Cruz, Rio de Janeiro. Pesquisador do Núcleo de Ecologia Humana e Saúde do Departamento de Sociologia e Ciência Política, Universidade Federal de Santa Catarina, Florianópolis. **Visiting Professor, Department of Human Ecology, Vrije Universiteit Brussel**, Belgium. Endereço eletrônico: favila@matrix.com.br.

Introduction

Human health ecology is an expanding field of interdisciplinary research, combining data, drawn from distinct areas of knowledge. Human ecology is essentially interdisciplinary. The search for the origin and distribution of diseases demands a combination of theories and methods from both natural and social sciences. Epidemiological analyses must be based upon a sound knowledge of social interactions and institutions. The mere manipulation of mathematical data is an insufficient basis for health research. Statistical manipulation is not a substitute for reasoning based upon biological and social congruence.

Epidemiologists search for the origin and distribution of diseases. Observations and experiments must be repeated in order to permit natural phenomena to be fully understood and explained. Science is explanation achieved in accordance to certain rules. One of them is that explanation must be confirmed by further observations or experiments performed by independent scientists, following the same methods and procedures as detailed by the first author who published them. As there is a limit to repeated observations and experiments: hence the importance of statistics.

In science, as in daily life, we deal with samples, models, and images of a complex reality. As Plato said in *Republic,*

> Now do you suppose that if a person were able to make the original as well as the image, he would seriously devote himself to the image-making branch? Would he allow imitation to be the ruling principle of his life, as if he had nothing higher in him? (op. cit, p. 25)

But Umberto Eco (1990, p. 265) cautions us by saying that: "As the man said, for every complex problem there's a simple solution, and it's wrong."

Statistical analyses are usually based upon selected samples and samples must be representative of the whole. Taxonomists describe species studying samples of populations and these must contain specimens of different phases or stages of development, sexes, and forms, in polymorphic taxa.

Questionnaires and interviews are widely used when dealing with human subjects, but they must be conducted with carefully selected samples to provide valid answers to our queries. Enquiries made with patients in hospitals, for instance, cannot be generalized for the whole population, although this truism is often disregarded. Claude Bernard (1957, p. 134-135) had already condemned, in 1865, certain meaningless and bizarre calculations: "A startling instance of this kind was invented by a physiologist who took urine from a railroad station urinal where people of all nations passed, and who believed he could thus present an analysis of average European urine!"

Irrelevant, erroneous, false, and misleading, information based on flawed interpretation of statistics has always plagued scientific publications. In some cases, not only professional ethics has been breached, but risks to well-being and health has been fostered (HUFF, 1954).

In this article I will not deal with fraud, but with *bona fide* errors of judgment and the use of erroneous statistical concepts.

1. Categories and correlations

Faulty statistical correlations are a source of misunderstandings leading eventually to false conclusions. Susser's (1973) book remains one of the best introductions to causal thinking in the health sciences. Both Susser (1973) and Peller (1967) are good sources for instances of numerical correlations which are nothing more than the result of coincidence or chance. It is rather a sort of numerology. The historical records of cases of human plague recorded by the former Brazilian Plague Service shows a remarkable coincidence between the number of cases per month with the number of letters in names of the same months.

For a correct interpretation of statistical data, one must be thoroughly familiar with the several primary and secondary factors intervening in the disease or condition under scrutiny (Cross, 1983; Schad *et al*, 1983).

The order or sequence of events is often unknown or is mistakenly interpreted. Peller quotes an editorial in the New England Journal of Medicine (1961) about a child who after seeing several crippled peddlers selling pencils in the street, asked his father why the selling of pencils caused men to lose their legs. There is actually a correlation here, but the cause and effect were reversed in the child's mind. The information that a person died around three hours after eating is misleading. The catch is that everything that happens to us occurs approximately three hours after eating. Searching for a correlation with food poisoning is bound to lead nowhere.

In the medical literature we find a number of examples of correlations based upon notoriously imprecise data, or difficult - if not impossible - to measure, as those involving "Mediterranean diet" (Meddiet, 2004), stress (Selye, 1936), depression (Sharp; Lipsky, 2002), satisfaction (Newgarten, 1961), and pain (Woodrow *et al.*, 1972). Factors listed in tables for the evaluation of degrees of stress are at best imprecise and subjective (Selye, 1950). Pain is a very personal experience and "1 to 10" scales used with different patients are not comparable. Zola (1961) demonstrated that

> Pain is an especially interesting phenomenon since there is considerable evidence that its tolerance and perception are not purely physiological responses and do not necessarily reflect the degree of objective discomfort induced by a particular disorder or experimental procedure.

Indeed Zola observed that outright demonstration of distress depended on nationality or culture, the Irish being far more stoic as regards to pain than Italians during medical consultations.

My objective in this article is to caution against a common practice adopted by researchers using qualitative as well as quantitative methods of analysis in order to establish demographic categories. Such categories must be established having in mind exactly what we want to learn. Categories which are valid for one type of investigation are not necessarily applicable to other objectives, as we shall see. This statement may appear evident but actually this is

not the case, as I will endeavor to show. A good discussion concerning the categories of analysis was presented by Goldberg (1982).

1.1 Age

It is not easy to establish age-classes in a population sample when the exact age of individuals cannot be accurately determined. Ethnographers studying so-called primitive societies are well aware of that. The literature abounds with examples of errors in the dates of birth and death even when engraved in tombstones. Marrou (1975) gives a number of interesting examples when historians have been led to error in this respect. History students are advised to doubt, and to look for independent confirmation even when dealing with official birth certificates. During WW2 Europeans from several nations had good reason to falsify their identity papers, changing names and age. Legends of longevity of individuals from remote places disregard the fact that no birth records are available to confirm the alleged longevity. Where a tax is charged for birth certificates, poor people tend to falsify the actual date of birth to avoid legal fines - or to forget the exact date. When the researcher's objective demands the establishment of precise age classes, he must be very careful when checking his sources. Age class-related diseases are perhaps the easiest to correlate. Biological age is definite factor of disease, and there is no need to elaborate on this question.

1.2 Sex

This category is undoubtedly one of the most peculiar in behavioral and in epidemiological studies as disease transmission depends upon both, biological as well as behavioral factors. Sex related diseases must be separated from those dependent on sexual behavior. Epidemiologists tend to forget that sexual practices are independent from biological or phenotypic sex and gender. The epidemiology of AIDS or hepatitis B offers good examples. Exposure to risk depends on social comportmental mores and on

the behavior of individuals. Unless we take into account the multiple variations of coupling or pairing, we will be unable to explain the patterns of distribution of sexually transmitted diseases in a population. Instead of two columns (males/females) we need at least four: heterosexuals, bisexuals, homosexuals, the category "homosexual" being further split into subcategories, as female homosexualism involves fewer risks than male homosexualism. Is it difficult to get candid answers? So it is with declared age (21+ for example), diet, drinking, and financial status.

1.3 Marital status

If a sociologist or an ethnographer wants to find out how many couples in a population have appeared before a Judge of Peace and hold a legal marriage certificate, the category married--or-single is a valid one. For clinicians and epidemiologists it is meaningless. When a gynecologist asks a patient: "single or married?" he learns nothing that may be useful for diagnostic or medical purposes. As with reference to sex, what he needs to know is the sexual habits of his patient, the onset of sexual life, frequency of sexual relations, number of partners, and other significant details that may affect health. A woman or a man may be married and virgin and this is a valid cause for the annulations of marriage among Catholics. If not exceptionally common nowadays and in most societies, it was rather usual among couples in past centuries where it was customary for an adult husband to marry an infant or pubescent girl and delay having sexual relations until puberty. Or else, a patient of either gender may be legally single and lead a regular sex life, with one or several partners.

1.4 Literacy and education

Educational standards differ widely among countries and among schools in the same country, from kindergarten to graduate school. Grading systems are different throughout the world. Ranking of colleges and universities show how little we may be

able to infer from a diploma or certificate of studies. (Blane, 1985; Stronks *et al*, 1996). These facts are well publicized, but this category appears nevertheless to have general acceptance in epidemiological studies as a blank check.

Recent surveys made by national and international agencies show clear differences in the proficiency of Brazilian students from public and private schools. Contrary to what happens in England, for instance, public schools in Brazil, with a few honorable exceptions, rates lower than private ones in teaching standards.

1.5 Occupation

Early correlations between occupation and health were established in 1848 by Rudolf Virchow, then a young medical doctor commissioned by the Prussian government to investigate an epidemics of typhus among coal miners in Silesia.

Occupational disease is an important issue, as practically all kinds of jobs pose some sort of risk to health. Actuaries working for health insurance corporations have been described as experts in turning the uncertainty of life into the certainty of profit, using simply a formula, where occupation is one item of the equation.

We must be aware that for epidemiological or statistical purposes, job descriptions *per se* may be misleading. Actually, how a job is performed is the factor that exposes individuals to risk. A person may belong to a group of risk without exposing itself to risk.

1.6 Religion and cult

An important category, it includes rituals, fasting, sacrifices, deprivations, pilgrimages to sacred or holly places, which entail risks to health. Rat temples in India offer a good example. Pilgrimages to Meca and to Virgem da Lapa in Brazil, Lourdes, and hundreds of religious shrines across the globe also involve risks. Including those linked to protests, intolerance, and terrorism.

Crowding exposes individuals to risks, from environmental contamination to direct transmission of pathogens. A recent example is a bacterial infection known as legionellosis. It has been named Legionnaire's disease after an outbreak, which victimized people attending a meeting of the American Legion in 1976 in Philadelphia.

Conversely, during the plague epidemics that ravaged Europe in the Middle Ages, monks and nuns living isolated in monasteries were spared, due to the better sanitary conditions of their living environment.

One must be cautioned against the acceptance of false anachronical explanations, as the proscription of pork meat by Muslims and Jews, as being related to the prevention of parasitic diseases.

In an epidemiological investigation it is important to be aware of the distinction between "declared religion, cult, or persuasion" and "religious regular adherence". This question is not always answered in a regular interview or in a written questionnaire - but it is of fundamental significance. Religious adherence must not be taken for granted, or we would not detect delirium tremens among Muslims, BSE (Bovine Spongiform Encephalopathy) in hindus, or pork meat triquinelosis in orthodox jews.

1.7 Race, nationality, ethnicity, and culture

This is another confounding variable, as we shall see (BRADBY, 1995; SMAGE, 1996; TEMPLETON, 1998).

Stephen Morse (1995, p.294) criticized medical doctors who ignored the meaning of certain concepts in epidemiology: "It is embarrassing to see how many members of today's medical community do not know the definitions for race, ethnicity, or culture."

Biologists agree that the species *Homo sapiens* did not split into distinct races (Dunn, 1961). Epidemiologists that use "race", "ethnicity", "nationality" and/or "culture" as a category, usually get nowhere. In general, race is defined by skin color but in the Uni-

ted States, ancestry is also taken into account. As a reproductive isolating mechanism, blood group Rh is far more effective than skin color. The euphemism "Afro-American" is also meaningless; as people from the African continent come in many shades of color, belong to different ethnical groups and to many nationalities. The same applies to "Hispanics" as Morse showed: a pure Japanese national, born and raised in Peru, where he was elected President is "Hispanic".

Among the many examples of confusion, Donald Cooper (1986, p. 486) who described the cholera epidemic that hit Belém in 1855 concluded that "No class or race escaped the ravages of cholera completely, but black people paid by far the highest tribute to the disease. It seems certain that no less than two-thirds of the victims of cholera in Brazil were blacks". Which was only to be expected? Black people suffered more not for being black or due to racial characteristics as claimed, but for being slaves living in sanitary conditions that exposed them to a higher risk.

2. Watch out for confounding variables

Association of variables must take into account distortions due to co-variation. Susser (1973) remarks that "In epidemiology, perhaps the commoner distorter variable is age". In fact, increasing age is positively associated with many other variables, and must be controlled to avoid spurious association. In a given example, rates of depressive psychosis in an English city appeared to be higher in married couples than in single individuals. When the age variable was controlled, it became clear that increasing age was positively associated both with marriage and with depressive psychosis. Older people were more likely to be married and were at greater risk of psychosis. We know for sure the date of the wedding, but not the onset of depression. Actually the rates of depressive psychosis were higher among the single and marriage was negatively associated with that condition.

"Race" may be linked to social conditions that expose people to malnutrition, unsanitary living environment, and restricted

access to medical resources.

Occupation and stress are often correlated. The difficulty is the imprecision in measuring stress. No single event is stressful for everybody, even for the same individual at different times and in different circumstances.

Conclusions

We must bear in mind Peller's advice that in some phenomena purely biological factors are causally involved; in others, both social and biological factors (PELLER, 1967).

As a consequence, we must resort to a combination of methods used by distinct disciplines (MURRAY *et al*., 2002). There is no single, universal, all-embracing methodology. The best we can use when we tackle problems that involves exact, natural, and social sciences is the equivalent to the Swiss army knife: a collection of tools in the hands of someone proficient in the choice and use of each tool, as required by each situation.

Epidemiology cannot be reduced to the mere application of formulae and calculations. It requires first and foremost the exercise of logical thinking, the proposal of hypotheses and logical and acceptable justification of the results.

Recebido em: 08.03.2011
Aprovado em: 5.06.2011

References

BERNARD, C. **An introduction to the study of experimental medicine**. New York: Dover, 1957.

BLANE, D. An assessment of the Black Report's 'explanations of health inequalities'. **Sociology of Health & Illness**, vol. 7, n. 3, p. 423-445, 1985.

BRADBY, H. Ethnicity: not a black and white issue: a research note. **Sociology of Health & Illness**, vol.17, n. 3, p. 405-417, 1995.

COOPER, D. The new "black death": cholera in Brazil, 1855-1856. **Social Science and Hist**., vol. 10, n. 4, p. 467-488, 1986.

CROLL, N. Human behavior, parasites and infectious diseases. *In* CROLL, N. and CROSS, J. **Human ecology and infectious diseases**, p.1-20. New York: Academic Press, 1983.

DUNN, L. **Race and biology**. Paris: UNESCO, 1961.

ECO, U. **Foucault's pendulum**. New York: Balantine, 1990.

GOLDBERG, M. Cet obscure objet de l'épidémiologie. **Sciences Sociales et Santé** 1:55-109, 1982.

HUFF, D. **How to lie with statistics**. London: Penguin, 1991.

MARROU, H-I. **De la connaissance historique**. Paris, Seuil, 1975.

MEDDIET. **Minutes of the meeting on the definition of the traditional Mediterranean diet in the context of the MEDDIET 2004** International Conference 21-22/04/2004, Athens, Greece, 2004.

MORSE, S. Tuberculosis susceptibility: race, ethnicity, nationality. Editorial. **New England Journal of Medicine**, Aug. 10, 1961. Available: http://www.promedmail.org. Accessed 18 July 2011.

NEWGARTEN, B.; HAVIGHURST, R. and TOBIN, S. The measurement of life satisfaction. **Journal of Gerontology**, *v.* 16, n. 2, p. 134-143, *1961*.

MURRAY, T.; KAY, J.; WALTNER-TOEWS, D. ; RÁEZ-LUNA, E. Linking human and ecosystem health on the Amazon frontier. *In* AGUIRRE, A.; OSTFELD, R.; TABOR, G.; HOUSE, C. and PEARL, M. (eds.). **Conservation medicine**: **ecological health in practice**, p. 297-309. Oxford: Oxford University Press, 2002.

PELLER, S. **Quantitative research in human biology and medicine**. Bristol: John Wright & Sons, 1967.

PLATO. **Republic**. The Internet Classics Archive.

Available: http://classics.mit.edu//Plato/republic.html Accessed 2 July 2011. p.25.

SCHAD, G. Human ecology and the distribution and abundance of hookworm populations. *In* CROLL, N. and CROSS, J. **Human ecology and infectious diseases**, p.187-223. New York: Academic Press, 1983.

SELYE, H. **Stress**. Montreal: Acta, 1950.

HARP, L, and LIPSKY, M. Screening for Depression across the lifespan: A review of measures for use in primary care settings. **American Family Physician**, v. 66 n. 6, p. 1001-1008, 2002.

SMAGE, C. The ethnic patterning of health: new directions for theory and research. **Sociology of Health & Illness**, v.18, n. 2, p.139-171, 1996.

STRONKS, K.; VAN DE MHEEN, H.; LOOMAN, C. and MACKENBACH, J. Behavioural and structural factors in the explanation of socio-economic inequalities in health: an empirical analysis. **Sociology of Health & Illness**, vol. 18, n. 5, p. 654-674, 1995.

SUSSER, M. **Causal thinking in the health sciences**. New York: Oxford University Press, 1973.

TEMPLETON, A. Human races: a genetic and evolutionary perspective. **American Anthropologist**, v.100, n. 3, p. 632-650, 1998.

WOODROW, K.; FRIEDMAN, G.; SIEGELAUB, A. and COLLEN, M. Pain tolerance: differences according to age, sex and race. **Psychosomatic Medicine**, v. 34, n. 6, p. 548-556, 1972.

ZOLA, I. Culture and symptoms - An analysis of patient's presenting complaints. **American Sociological Review**, v. 31, n.5, p. 615-630, 1966.

Resumo

Ecologia humana, análises estatísticas e a lógica das correlações válidas

Ecologia da saúde humana constitui uma das áreas de pesquisa interdisciplinar de maior expansão atualmente, combinando conhecimentos provindos de distintos campos do conhecimento.

Ecologia humana é essencialmente interdisciplinar. A investigação da origem e distribuição das doenças e agravos exige a integração de teorias e métodos tanto das ciências naturais como das ciências humanas. Há um limite para a replicação e repetição de observações e experimentos na pesquisa científica: daí a importância da estatística. Como instrumento, análises estatísticas baseiam-se geralmente em amostras selecionadas e as amostras devem ser representativas do todo. Informações irrelevantes, errôneas, falsas e enganosas, baseadas em interpretações viciadas de dados estatísticos são encontradas, com frequência, em publicações científicas. Em certos casos, não só a ética profissional é violada, mas riscos à saúde e bem estar são resultantes. Neste artigo não trato de fraudes, mas de erros de julgamento cometidos em boa fé e da utilização de conceitos errôneos utilizados em análises estatísticas. Meu objetivo é o de precaver contra práticas correntes em pesquisas que se utilizam de métodos qualitativos e quantitativos de análise, tendo em vista estabelecer categorias demográficas.

Palavras-chave: ecologia humana, correlações, análise estatística, categorias, variáveis de confusão.