DIVERSITY STUDIES ON MORTALITY DATA

by

J. IZSÁK* – P. JUHÁSZ-NAGY**

Received 19.04.1982.

Abstract

Nowadays the more and more accumulated morbidity and mortality statistics and the computerized health data banks make it possible and necessary to work out new methods for epidemiological investigations. This article raises the idea that the totality of the diseases appearing in any population might be subjected to investigation without the separate investigation of the individual diseases. The situation might be compared to that of the floristics and plant sociology. A simple example for the study of the totality of the diseases could be the investigation of the regularity of the distribution among the categories, for wich we use the relative diversity. This index has become quite wide-spread in the last decades in biosociological and ecological investigations. Distributions with characteristically small diversities had been investigated earlier in econometry and in statistical linguistics (see: Pareto-distribution and Zipf's Law).

The opportunities offered by the diversity studies are demonstrated by some USA mortality statistics. In the majority of the disease sections the diversity of the distribution of the deaths among the ICD categories of causes of death culminates around 15 – 35 years of age, after that it decreases. The diversity of the female groups is mostly larger than that of the male groups. On the graph curves made from the diversity indices, one could observe certain differences among the white, the negro and other races. Here one should consider the differences in life-style (way of life). In the case of some sections, the diversity curves of men show a characteristic hollow around 50 – 60 years of age, which could be explained by special civilizational effects. Both diagnostical and biological factors take part in the formation of the character of the diversity curves. For a more detailed study further investigations and considerations are needed.

Key – Words: Disease Pattern of Population, Diversity, Entropy, Mortality Data, Ageing, Sex Difference.

* Teachers' College, Szombathely H – 9701, Dep. of Biology
** Eötvös University, Department of Plant Taxonomy and Ecology, Budapest

Introduction

In general, the chief aim of preparing comprehensive morbidity and mortality statistics is to offer in a given circle in specified detail a survey of the rate and totality of the listed diseases and causes of death. Instead of taking hold of the group of data together, however, the mortality and morbidity rates are studied separately, without excluding, naturally, their comparison. Generally, the view that would consider every disease as one of the many without discussing individual morbidity rates and that would turn attention to the characteristics of the "pattern of diseases" of the studied population does not gain ground. Although exploring certain interconnections is possible evidently only if a wider circle of phenomena is comprehended, just as floristic and faunistic research cannot substitute the exploration of biosociological and ecological connections.

The question arises, naturally, wether the totality of diseases and causes of death in a population can be considered as a natural object of investigation, as the flora and fauna of a biotopological unit can. Most of the diseases and causes of death themselves are not so well definable entities as the species of the living world ([2], [5], [10], [14]). Still, it is possible that a more comprehensive study of the great number of morbidity and especially that of mortality statistics at our disposal leads to the discovery of new phenomena [11].

One of the reasons of the certain disinterest in the collections of health statistics is that they are not sufficiently informative for the investigators in the field, while the comprehensive studies and methods are rather neglected [3]. Only the investigation of health status index, used as a global index-number of the health status of the population can be mentioned here. In developing its different variety, however, mainly health-organizational, economical and sociological aspects took a prominent part [1].

The *index of diversity*, that has been used for the analysis of floristic and faunistic statistics for a long time, can also be suitable for the study of the morbidity data in accordance with the above criteria.

Since the 1920's G a r t h s i d e, C o r b e t, W i l l i a m s and others have been reporting a characteristic feature of the entomological material collected by them: it was striking that a great number of the collected individuals belonged to only a few species, while most of the represented species was represented by only a few or only a single individual ([21], [17]). It was observed on the entomological, ornitological and other materials collected under various circumstances that ranking the frequencies, the frequency of the species decreases rapidly and in a characteristic way. For the analysis of this characteristic distribution, starting off from binomial coefficients, sequences as probability distributions were used ([21], [16]).

It has to be mentioned that similar distribution has been described concerning word-frequencies within a given text [22], furthermore, this feature can be observed in the case of Pareto-distribution discussed frequently in econometrics. In the statistical discussion of the special litera-

ture the direction of investigation goes back to Lotka (see de Solla Price [18]).

Returning to biological aspects, beside extensive faunistic and floristic material, similar questions have been raised beforehand in the investigation of parasitological statistics ([21], chapter 8).

Diversity is widely studied today in ecological and biospheric research. It was Herdan in 1957 who was one of the first ones to apply the method for morbidity statistics [6]. Beside investigating the distribution of diagnosis of hospitalized patients in England and Wales he was also dealing with the measurement of diversity. In his pioneer-like paper he took up the position that the nature of the frequency distributions of diagnoses stands between the frequency distribution of words and that of taxons in biology. In accordance with his view he considers the numerical characteristics of diagnosis distribution adequate for the investigation of different diagnosis practices.

The interesting studies of Herdan, however, were not continued. E. g., recently, Höpker [8] touched upon the method of ranking diagnosis frequencies, referring just to Zipf's linguistic observations, but he does not even mention Herdan's paper.

As it has been indicated earlier, in our view, drawing a parallel between the analysis of the statistics of morbidity and mortality and that of floristics-faunistics comprises a lot of unexploited possibilities. For this reason, as a first step we started to study the index of diversity of the frequency distribution of morbidity and mortality.

An interesting phenomenon could be observed even at our first attempt: the index of diversity of hospitalized female patient groups was generally greater than that of the corresponding male groups [12].

A similar study was carried out on the Hungarian mortality statistics as prepared according to the "A" list of the International Code of Diseases (ICD); the same phenomenon could be observed, here too [13].

Recently we described the lognormality of frequency distributions of death cases on U. S. mortality data [13a]. (Distribution of the frequencies is an other side of the diversity problem).

Materials and methods

In further studies we planned to continue our investigations on a more detailed and more generally known statistical material, laying special emphasis on the sex differences in the index of diversity.

Material meeting these requirements can be found primarily in the field of mortality statistics. Finally, we decided upon the annual publication of Vital Statistics of the United States, the second volume of which contains mortality data [19]. The studied part is based on the three-digit list of ICD and has the following advantages:

a) usually the number of categories within the sections of ICD is sufficiently great

b) the frequency data is relatively large, as compared to similar statistics
c) the age groups contain the same number of years, not even older ages are combined
d) division according to race makes the introduction of a new aspect of investigation possible.

The categories of causes of death are indicated according to the 8th revision of ICD with some modifications that are unimportant for us.

Naturally, other mortality statistical data are also suitable for similar analysis.

The analysis was carried out only on sections I., II., III., IV., VII., IX., X. and XIV. The rest of the categories were excluded because of the small number of causes, or because the statistics of the category did not make it possible to compare diversity indices as a function of sex. Section E XVII. was excluded because of the special nature of its causes of health. A different human biological concept is feasible, however, which requires the *joint* consideration of *every* mortality category.

The categories to which several, quite different causes of death were assigned were excluded from the studied sections. E. g. the category "Other tuberculosis including late effects", that includes the tuberculosis of the urogenital system, tuberculosis of other organs and the late effects of tuberculosis (categories 016, 017, 019) was left out. The four digit subcategories were reduced into the three digit categories. Although the decision concerning exclusions and contractions contain subjective elements also, still, the inclusion of all of the mortality categories would not have been correct.

Accordingly, a certain index of diversity of deaths belonging to a given race, sex, age group and disease section, is connected to the distribution of causes of death within the categories. It is important to note that if a cause of death was left out, the corresponding deaths were also omitted.

For rough information some characteristics of the studied 1974 statistics are reported in the first two columns of Table 1.

Measuring the diversity

Several indices have been suggested as the measure of the diversity of class frequencies ([16], [7], [14]). From among the so called *Brillouin-index* is of central importance. Its formula is:

$$H = \frac{1}{N} \log \frac{N!}{\prod_{i=1}^{s} N_i!},$$

where N ($\neq 0$) is the number of elements in the universe, s is the number of subclasses, N_i (i = 1, 2, ... , s) is the number of elements in the i-th subclass.

Table I

Basic data related to the calculation of the diversity index and the m:f ratio according to the races

Section	Number of causes+	Death ratio++	*m:f* ratio			
			white	negro	other races	all together
I. Infective and parasitic diseases	45	79.5	10:11	5:16	4:9	8:14
II. Neoplasms	37* 41**	90.2	9:13	6:16	7:10	8:13
III. Endocrine, nutritional and metabolic diseases	7	84.8	4:17	7:11	3:4	4:18
IV. Diseases of the blood and blood-forming organs	–	–	–	–	–	–
V. Mental disorders	–	–	–	–	–	–
VI. Diseases of the nervous system and sense organs	6	56.3	11:10	8:11	3:2	10:11
VII. Diseases of the circulatory system	29	88.0	4:18	7:15	2:17	2:18
VIII. Diseases of the respiratory system	16	58.5	6:16	11:11	6:7	7:15
IX. Diseases of the digestive system	11	74.9	6:16	8:13	7:8	6:16
X. Diseases of the genitourinary system	5	54.3	9:13	5:16	5:5	8:14
XI. Complication of pregnancy, childbirth, and the puerperium	–	–	–	–	–	–
XII. Diseases of the skin and subcutaneous tissue	–	–	–	–	–	–
XIII. Diseases of the musculosceletal system and connect. tissue	–	–	–	–	–	–
XIV. Congenital anomalies	5	69.3	8:10	10:7	0:2	8:10
XV. Certain causes of mortality in early infancy	–	–	–	–	–	–
XVI. Symptoms and ill-defined conditions	–	–	–	–	–	–
E XVII. Accidents, poisonings, and violence	–	–	–	–	–	–

+number of considered causes (categories) of death ++death ratio of the considered categories in the sections%
* males ** females

Evidently, H can be considered as the approximation of entropy of the polynomial distribution with parameters s, p_1, p_2, ..., p_s from the sample, where the approximation of probability p_i is $\frac{N_i}{N}$. Since H depends upon the s number of subclasses, from many aspects the quotient

$$J = \frac{H}{H_{\max}}$$

is a more adequate index, where $H_{\max}$ is the maximal H value that can be reached in the case of s number of subclasses. Its formula is:

$$H_{\max} = \frac{1}{N} \log \frac{N!}{\left(\left[\frac{N}{s}\right]!\right)^{s-r}\left(\left(\left[\frac{N}{s}\right]+1\right)!\right)^{r}},$$

where [ ] is the sign of entier function and $r = N - s\left[\frac{N}{s}\right]$. In cases of large number of elements and pre-fixed number of categories the quantity

$$J' = \frac{-\sum_{i=1}^{s} \frac{N_i}{N} \log \frac{N_i}{N}}{\log s}$$

is a good approximation of J, in the numerator of which the so called *Shannon-index* can be found.

The above quantity J is called the *relative diversity* of the distribution. Henceforth J will be mentioned simply as the index of diversity.

The statistical problems of the different indices of diversity are dealt with by many authors ([16], [9]).

The index of diversity was calculated by a computer program in FORTRAN IV, according to the previously described formula. The logarithm of the factorials of numbers greater than 45 were approximated by Stirling's formula. It has to be noted that the value of J falls always between 0 and 1, and that the diversity for the degenerate case with $N = 1$ was not interpreted.

In order to sum up the calculated indices of diversity curves were drawn from the series of diversity indices belonging to the given disease section and race (Figs. 1, 2 and 3a).

We wanted to demonstrate as much curves as possible, because, in accordance with the aim of this paper, the informative role of the curves is great. In choosing the curves we attempted to demonstrate examples for as much characteristic features as possible. It is not possible, of course, to show all the curves, and those curves where the total rates for the different age groups were very small were not drawn at all.

The first point of the curve is not connected to the second, indicating that it belongs to the 0 year age group (infant mortality). The second point

is connected to the third, although the second point belongs to the 1 – 4 year age group, thus, as distinguished from the rest of the age groups, it includes only 4 years. Naturally, we could have given the index of diversity of the age group 0 – 4 (i. e. those below 5 years of age), but because of the special nature of infant mortality this age group could not be compared with the other groups at all. For this reason we found the present illustration more informative.

The curves of larger frequency data are naturally much smoother than the others because of the relatively small sampling fluctuation. On the other hand, in sections where relatively small amount of the total death cases were included in the investigation (sections VI., VIII. and X., see Table 1), one has to be careful in interpreting the indices of diversity and the curves. The small number of categories in sections III., VI., X. and XIV. also warns us to be careful.

Depending on one's point of view one could emphasize different characteristics of the curves, thus, we cannot enter into all the details of our analysis. The following facts are worthwile to mention by all means: – The maximal element of the diversity index is mostly between the 15th and 35th years of life. E. g. see the curves in Fig. 1. and curves *e)* in Fig. 2.

– The indices of diversity in the middle age-groups are greater for females than for the corresponding male groups, almost without exception.

– After reaching a maximum generally there is a continuously decreasing tendency (e. g. see most of the curves in Figs. 1. and 2.). To this only sections VI. and X. (diseases of the nervous system and sense organs, and diseases of the urogenital system) are exceptions where the curves run differently (Fig. 2., curves *a)* and *c)*). In sections II., VII. and IX. (neoplasms, deseases of the circulatory system, diseases of the digestive system) the second rising part (Fig. 1., curves *c)* and *e)*, and Fig. 2., curves *e)*) can be traced back to the cessation of the cause eliciting the previous marked decrease.

The occuring characteristic hollow is more marked in the male group, and is the most prominent in the white male group in section VII. (disease of the circulatory system, Fig. 1., curves *e)*). A similar tendency can be seen in section VIII. (diseases of the respiratory system) in the white race also (Fig. 2., curves *f)*).

Since the difference between the corresponding curves of the different races was unambiguous and trivial, in order to increase the number of elements the diversity indices were calculated after combining the data of the different races. At this point, naturally, the number of outstanding peaks due to sampling fluctuation decreased. The curves are very similar to those of the white race. The reason for this is that the role of the distribution of the groups with large number of elements plays is greater than the arithmetical ratio in forming the diversity index of the combined groups (!) The curves received after combining the races are shown for sections III., VI., X. and XIV. in curves *a)* – *d)* in Fig. 2.

Finally, the sections were also combined, leaving the breaking down by race and sex. In this case the number of categories of causes of death

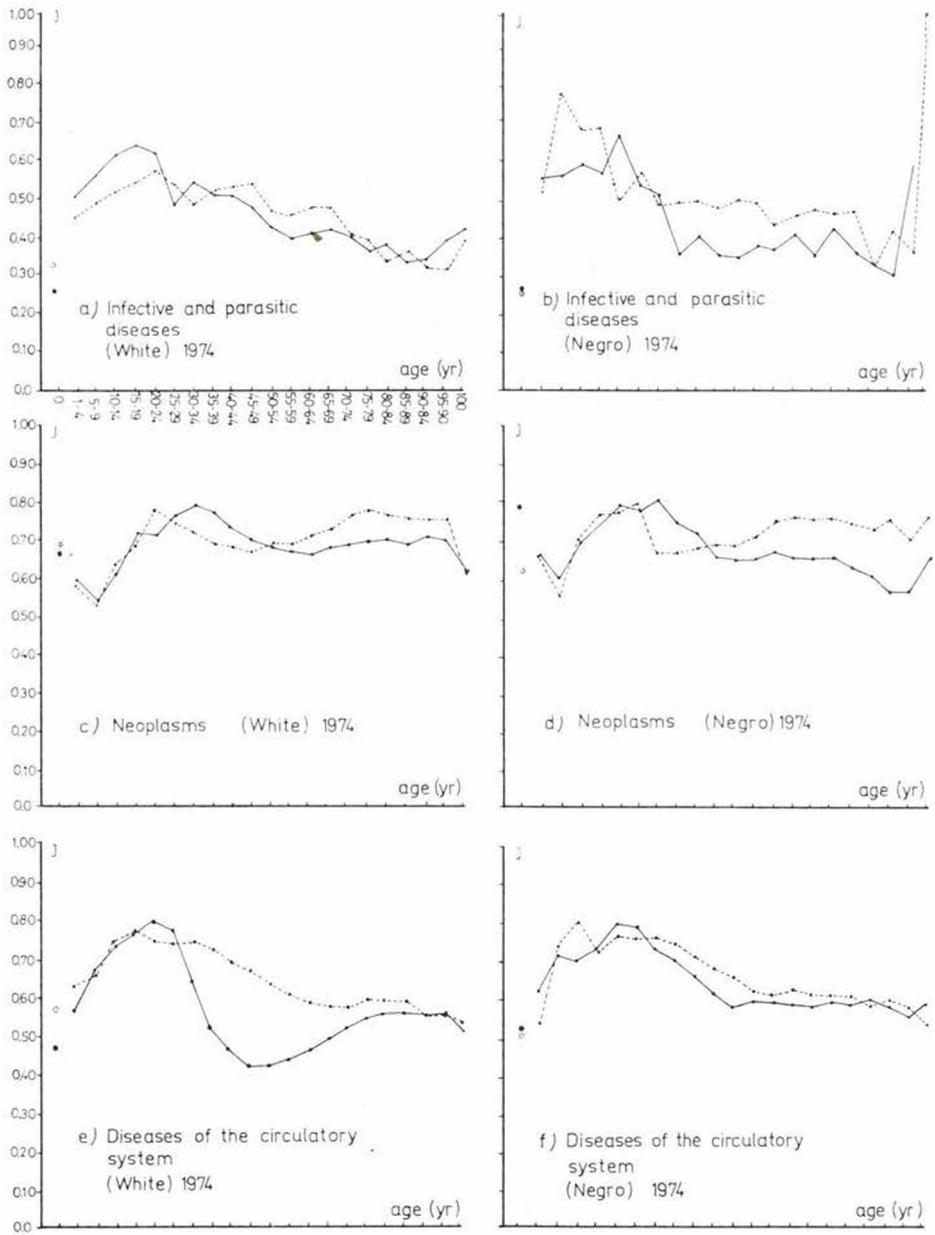

Fig. 1. Solid line and ●: males, broken line and o: females. The maximum point usually falls between 15 and 24 years of age. After reaching a maximum the index of diversity usually shows a monotonously decreasing tendency. The curve of males in *e)* is an exception, where the deep hollow is definitely followed by a rising period. It is generally outstanding that in the middle age-groups the diversity of female deaths is greater. In comparing the races there is a minimal deviation in the *c)* and *d)* pair of curves, while there is a marked deviation in the *e)* and *f)*pair of curves in the male group.

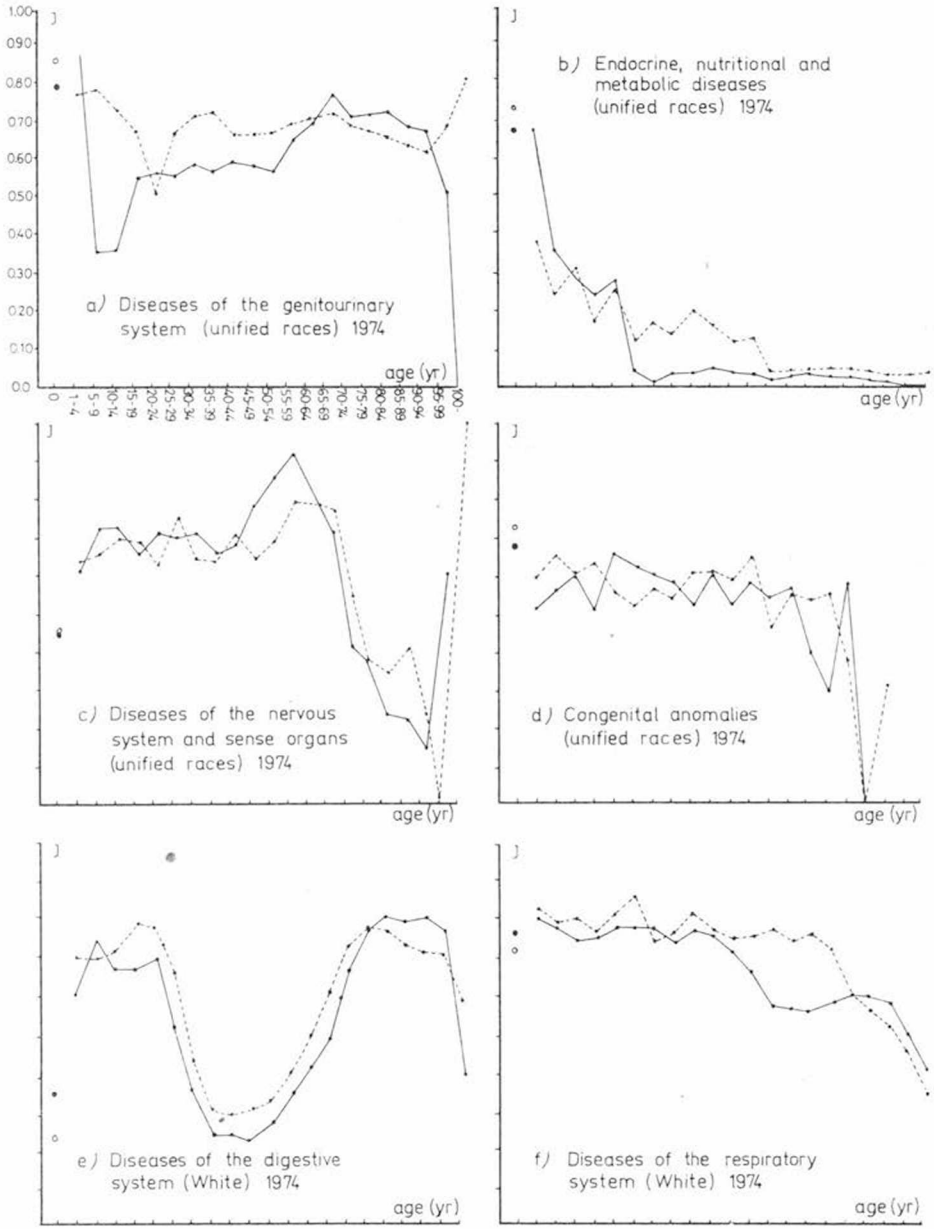

Fig. 2. Solid line and ● :males, broken line and o: females. The information concerning the curves of Fig. 1. are partially valid for curves *e)* and *f)*. The shape of curves *a)*, *b)* and *c)* differ markedly from that of the rest of the curves and from those of Fig. 1. The very small diversity in the older age groups in curve *b)* is conspicious.

as a sum of the number of section categories is 161 for males and 165 for females. The arising diversity index series are depicted in curves *b)* – *d)* in Fig. 3. It has to be remembered, here too, that the sections containing large number of elements (neoplasms and diseases of the circulatory system) dominate in forming the index of diversity.

The relatively smooth run of the curves show that sampling fluctuation is quite small, especially in the middle age-groups.

It can be pointed out that the maximal value observable in the 15 – 25 year age group both in males and females is the smallest in the white race and is the largest in the group "other races". The location of the maximal value behaves differently: in both sexes it occurs the earliest in the group "other races", and the latest in the white race. It is noteworthy that the

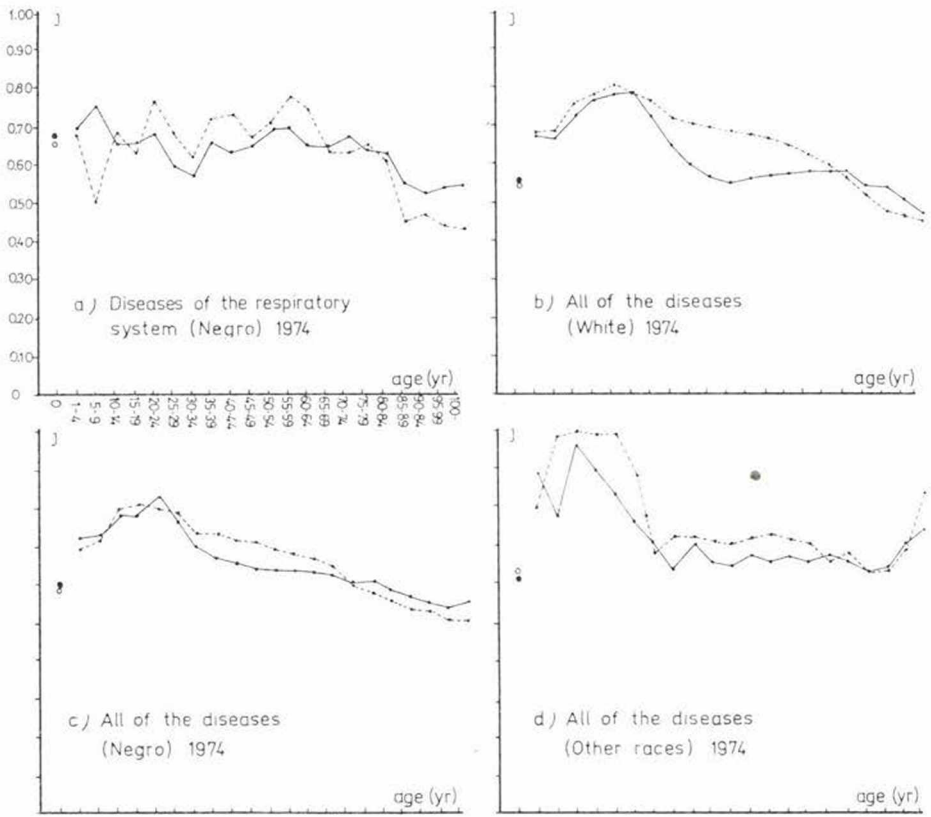

Fig. 3. Solid line and ●: males, broken line and o: females. After combining the sections a smoother curve is obtained (curves *b)*, *c)* and *d)*). Because of the greatest number of elements curve *b)* is the smoothest in both sexes. A maximum of diversity around 15 – 24 years of age, the subsequent even decrease and the female dominance in the diversity values in the middle age-groups can be observed. The hollow described in the text is shown in the curves of white males, and to a lesser extent in that of black males.

pairs of curves in the group "other races" run similarly to each other, but differ from the four other ones (Fig. 3., curves *d)*).

The above mentioned hollow can be observed after combining the sections in the group of white males, and to a lesser extent in the group of black males (Fig. 3., curves *b)* and *c)*).

It has to be emphasized in particular that the sample size was increased in two completely different ways. In pooling the races the number of cases is increased, while in combining the sections the number of categories is increased. The first type of sample size increase usually does not cause great difficulties, but the increase of the number of categories usually encounters difficulties. Similarly, the question of sampling fluctuation is also open to two interpretations.

It is not easy to explain the tendency of changes of the index of diversity as a function of age. Beside biological factors the problems of mortality statistics that have been discussed in many places have to be considered. Even if we accepted the set of the mortality categories, the classification of certain causes of death would still come up as a basic problem. It is well known, that the classification of the ICD lists was based on an eclectic system of aspects. E. g. the causes of death in the sections of infective diseases and neoplasms could be divided according to organs, or influenza could be listed among the infective diseases, etc. It has to be considered also, that frequently, there is a long interval between the onset of disease and death itself.

Notwithstanding, we think that basically, there is a biological background behind the characteristic run of the curves. This is supported by the fact that after combining the sections, as described previously, when the separation by sections no longer exists, the curves become very similar to the curves of most of the sections. And the other way around, the appearance of the characteristic features in several of the sections indicate the common and general nature of the underlying phenomena.

The hollow in the curves that has been mentioned several times can be explained by the fact that the frequency of some diseases peculiar to civilized communities (e. g. cardiac infarction, lung cancer, diabetes, hepatic cirrhosis) increases abruptly in the affected age groups. The greater concern of males also speaks for the role of the way of living. Because of their special nature the effects of civilization appear in narrow bands and because the resistance against them is very low, the distribution of the causes of death becomes less even, wich in turn decreases the index of diversity.

The development of the maximal value of the series of the index of diversity can be traced back to the fact that "aimed" selection falls into the background by the age of sexual maturity, after about 30 years of age the biological functions begin to break down in a more deterministic way. The phenomenon should be studied further for a more detailed explanation.

Sex differences in the index of diversity

Certain sex differences can be observed in the curves of Fig. 1 – 3. The index of diversity is usually greater in the middle age-groups of females than in the corresponding male groups. This can be explained primarily by the fact the above discussed hollow affects the male groups in a greater extent. Still, it cannot be stated that the female dominance described previously ([12], [13]) can be traced back to the above, clear-cut phenomenon. More detailed analysis reveals that – as opposed to the hollows in the curves – female dominance appears in practically every section and in most age groups.

For the sake of easier survey let us mark with the letter m (or f) the event that the index of diversity of the male (or female) group is greater. E. g. in section I. in the row "white race" the sign of the first event is f, because within this section the index of diversity of the white female infants is 0.310, while that of the corresponding male group is smaller, namely 0.270.

Undoubtedly, information is lost by not considering the degree of difference, but the female dominance in the index of diversity can be demonstrated this way, too. It is not calculated either, whether the difference in the indices is significant. Because of the large number of the diversity index pairs, however, it is improbable that the arising bias changed the pattern.

In the last four columns of Table 1. the $m:f$ ratios are given by sections and races and after the unification of races. The sum of the m and f events in the $m:f$ ratio is mostly less than 22, which is the number of the age groups, because in the majority of the age groups the two indices of diversity are equal or one of the members of the pair of groups contains only one element, thus its index of diversity and the comparison itself are meaningless. It can be seen that by races and sections the index of diversity was greater in the male group only in 3 cases of the 27. After the unification of the races there is a female dominance in all of the 9 sections. This is not explained solely by the hollow in the curves.

Concerning the index of diversity as a function of age it can be noted that the female dominance is most marked in the middle age-groups, although it is quite frequent in the rest of the age groups also. The different extent of the hollow as a function of sex may play an important role in this.

To sum it up, the sex differences in the index of diversity can be traced back partially to the relatively low index of diversity in the middle-age groups of males, eventually to the differences in the way of living. Sex differences do appear, however, independently of these causes, too.

It deserves attention that the sex difference is distributed among the sections. When the sections are considered separately, especially in the middle age groups, the phenomenon can be traced back to some causes of death widely known to be more frequent among males. Since female dominance can be observed in every studied section, the elementary explanations similar to those described above have their limits, and the solution

to the problem requires not the point of view of the specialist. The sex difference in the index of diversity has probably a more general, human biological background.

The questions of sampling fluctuation

Although a lot of authors have dealt with the sampling fluctuation of the index of diversity, in our case a direct observation seemed to be the most promising approach. For this reason the study on the data for 1974 was repeated on that for 1975 [20]. The basic components of the difference are random deviation ("error") and the difference arising from epidemiological fluctuations wich may manifest itself even if the number of cases is quite large.

The corresponding series of the diversity indices could be illustrated by curves again. The comparison of the curves of the two years showsn convincingly that the fluctuations due to small number of case may be quite large occasionally, but if the number of cases is large enough, good reproducibility allows for quite detailed analysis. Incidentally, after the first glance greater importance can be attributed to the race differences, too. The curves in Fig. 4. illustrate in the case of neoplasms that the race differences are greater than the difference between the two years.

The race differences are probably due primarily to those in the way of living.

The curves were analysed together in more detail from the aspect of differences, disregarding at this point the curves belonging to the combined groups. Thus, there were altogether 108 curves, since this is the number of free combinations of the years, sexes, races and disease sections. We considered it a basic question to what extent the above four parameters, as compared to each other, contribute to the shape of the curves.

A relatively simple statistical method was applied to investigate this question. To be clear, let us consider the following simple example. Let curve y be given and let us examine what parameter-change causes the greatest change in the shape of the curve. Deciding on some degree of alteration, the shape of the curve can be compared with that differing only in sex, then with that differing in year, with that differing in race (two), and finally with that differing in the 8 disease sections. In the latter two cases turning to the average of differences the degree of the effect of the four parameters can be measured by the above differences. The difference between two curves can be measured by the quantity

$$r'(x, y) = \sqrt{\frac{1}{22} \sum_{i=1}^{22} (x_i - y_i)^2} ,$$

where x_i and y_i ($i = 1, 2, \ldots, 22$) are the indices of diversity belonging to the i-th age group of curves x and y.

It has to be considered, however, that in some curves certain points (diversity indices) are missing, because the total number of cases is 0 or 1.

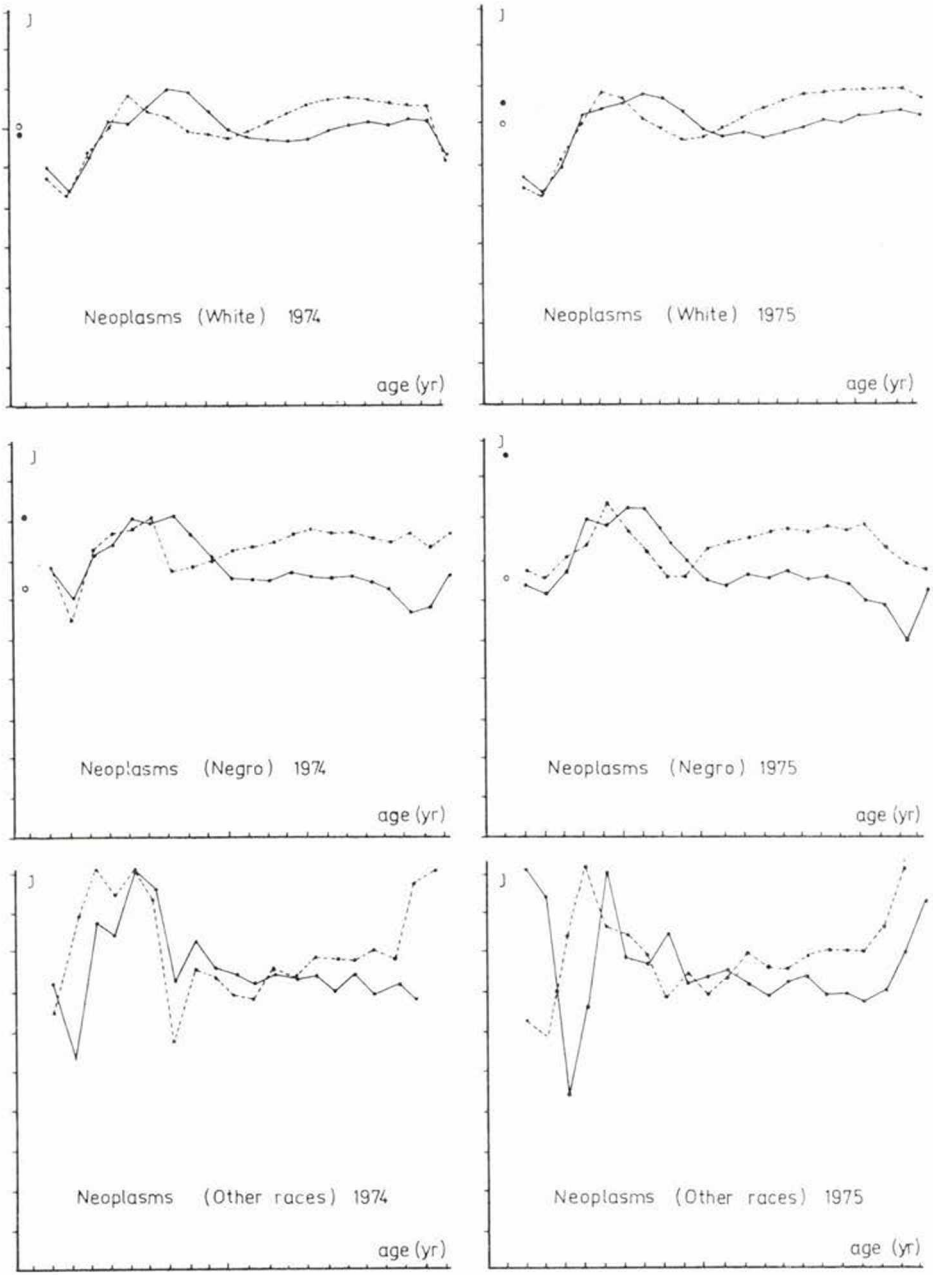

Fig. 4. Solid line and ●: males, broken line and o: females. The deviation due to race is visibly greater than the yearly differences.

If any point of the curve was missing, then the i-th coordinates were disregarded when comparing with all the other curves. At this point, however, it is more adequate to use quantity

$$r(x, y) = \sqrt{\frac{1}{22 - K} \sum_{i=1}^{22-K} (x_i - y_i)^2},$$

where K is the number of coordinate pairs in which at least one of the coordinates is meaningless.

From this simple example we arrive at our concrete problem so that g runs all over the set of the 108 curves, and the arising differences are summed up according to the parameters. As every pair of curves appear twice in this case, the sums are divided by 2. Thus, the following four quantities are arrived at as the measure of the forming effect of the parameters:

– average sum of deviation concerning year:

$$r_1(x, y) = \frac{1}{54} \sum_1 r(x, y)$$

– average sum of deviation concerning sex:

$$r_2(x, y) = \frac{1}{54} \sum_2 r(x, y)$$

– average sum of deviation concerning race:

$$r_3(x, y) = \frac{1}{108} \sum_3 r(x, y)$$

– average sum of deviation concerning disease section:

$$r_4(x, y) = \frac{1}{432} \sum_4 r(x, y),$$

where the sums are extended to curve pairs differing only in year, sex, race and disease section, respectively. (It has to be noted that the problem can be considered to that in an experimental situation with four variables in each of its plots one experiment is given, the result of wich, however, is vector-valued.)

The above sums were computed by a short computer program; the results were:

$$r_1 = 0.183$$
$$r_2 = 0.186$$
$$r_3 = 0.223$$
$$r_4 = 0.344$$

The results reflect that it is belonging to a disease section that determines primarily the shape of the curve, then come the effects of race, sex

and finally, but barely falling behind, the factor of year. It is probable that with greater number of cases quantity r_1, considered as the measure of random error, would not approach so closely quantity r_2, measuring the effect of sex. At any rate, the rank order of r_1, r_2, r_3 and r_4 turned out to be as expected.

In order to estimate the real degree of fluctuation due to the small number of cases the 1974 and 1975 diversity indices were compared with each other for the disease sections of the white males. Accordingly, it can be state that only if the number of cases is grater than about 8 – 10 thousand is the deviation of the diversity index less than the allowed 5 – 10%. It can also be stated that fluctuations that cannot be neglected do appear even with quite large number of cases. If $N < 100$, the indices of diversity are not reliable at all.

Summary

In the past few decades, as a result of different favourable phenomena more and more suitable statistical assemblies are available for epidemiological studies than before. As it is to be expected this will lead to more aspects and methods of epidemiological studies. In the present paper we attepted to offer an example for this possibility.

The investigation is based on the mortality statistics of Volume 2 of the health statistics annual Vital Statistics of the United States, 1974 and 1975. It can be stressed as a basic feature of the study that surpassing the discussion of the individual diseases (causes of death), the object of investigation constituted the mortality distribution of the population, or rather its certain characteristic feature that could be measured by numbers. It was found that studying the diversity or entropy of distribution is an appropriate way of discovering a new circle of phenomena.

The results were described in more detail by curves in wich the so called diversity index of the distribution of the causes of death was co-ordinated to the given age group of the deceased. The curves were prepared for groups classified according to race, disease sections and sex.

The problems arising from the ICD classification of diseases (causes of death), the very small number of categories within certain disease sections, the sometimes small number of elements in the sample can be pointed out from the factors hindering and restricting the investigation. The systematic bias of mortality statistics that has been described frequently has to be mentioned separately. Still, the characteristic run of the curves is worthy of attention.

In most of the nine studied sections the index of diversity reaches its maximum at about the age of 25 – 35 years, then it decreases continuously until the very old ages. In certain cases, for the middle age-groups, there is a hollow part of the curve wich may be traced back to the very high rate of certain causes of death peculiar to civilized communities. The phenomenon is especially expressed for white males. There is a distinct difference in the curves as a function of disease sections and sex; deviations as a function of race are smaller.

In agreement with our previous findings the diversity indices of the female groups were generally greater again.

In the course of studying sampling fluctuation it was established that fluctuation can be kept at a low level only if the total number of cases is large (about 8 – 10 thousand deaths).

A simple study was performed to decide to what extent the year (1974 and 1975), sex, race and disease section, as compared to each other, contribute to the shape of the curves. The deviations due to year were considered here as sampling fluctuations. It was found that belonging to a disease section is the decisive factor in determining the shape of the curves.

Returning to our investigation concept, we see quite some opportunity in the processing of the above mentioned morbidity and mortality data from a relatively new aspect. E.g. it would be worthwile to elucidate whether undernutrition, exhaustion and similar states cause smaller or greater variousness (diversity) in morbidity (or mortality).

The discovery of the background of the phenomenon also call for further investigations. At anv rate, quite general biological factors also have to be thought of. Reliability theory may be a discipline that might offer an appropriate frame for modelling and explaining the phenomenon.

If the supraindividual aspect is stressed, the „frequency role" of certain diseases (causes of death) becomes manifest, wich may be taken over by another disease as the given disease falls into the background, yielding the relative stability of the index of diversity.

Furthermore, it seems to be perspective to elaborate a line of investigation which considers a certain population and the corresponding diseases as a jointly moving pair of objects.

For the time being this is only fiction, but the rapid growth of morbidity and mortality statistics (primarily in human context), and the spreading of informatics press for starting similar investigations.

Key-Words: Disease Pattern of Population, Diversity, Entropy, Mortality Data, Ageing, Sex Difference

REFERENCES

[1] Balinsky, W., – Berger, R. 1975. A review of the research on general health status indexes. Med. Care **13**: 283 – 293.

[2] Doerr, W. – Jacob, W. – Nemetschek, Th. 1975. Über dem Begriff des Krankhaften in der Sicht des Pathologen. Internist **16**: 41 – 48.

[3] Erhardt, K. 1977. The underutilisation of vital statistics. Am. J. Public Health **67**: 325 – 326.

[4] Gavrilov, L. A. – Gavrilova, H. S. – Yagusinski, L. S. 1978. Basic pattern of ageing and death in animals from the standpoint of reliability theory. Zh. Obsh. Biol. **39**: 734 – 742. (In Russian, with an English summary)

[5] Gross, R. 1975. Der Krankheitsbegriff in der Sicht des Klinikers. Internist **16**: 49 – 52.

[6] Herdan, G. 1975. The mathematical relation between the number of diseases and the number of patients in a community. J. Roy. Stat. Soc., Ser. A **120**: 320 – 330.

[7] Hill, M. O. 1973. Diversity notation and its consequences. Ecology **54**: 427 – 432.

[8] Höpker, W. – W. 1975. Zur Quantifizierung selektionsbedingter Gruppenunterschiede in der Epidemiologie. Meth. Inform. Med **14**: 144 – 149.

[9] Hutcheson, K. 1970. A test for comparing diversities based on the Shannon Formula. J. Theor. Biol. **29**: 151 – 154.
[10] Immich, H. 1975. Der Krankheitsbegriff in der Sicht des Biostatistikers. Internist **16**: 53 – 55.
[11] Izsák, J. – Juhász-Nagy, P.: 1981 Studies on the pattern of diseases. Magyar Tudomány ("Hungarian Science") **27**: 39 – 43 (in Hungarian)
[12] Izsák, J. – Juhász-Nagy, P. 1979. Studies on the diversity of morbidity data series. Biológia **27**: 177 – 183. (in Hungarian, with an English summary)
[13] Izsák, J. – Juhász-Nagy, P. 1981. Investigation on diversity of Hungarian mortality statistics. Ann. Univ. Sci. Hung. (Budapest), Sectio Biol. **22 – 23**: 35 – 43.
[13a] Izsák, J. – Juhász-Nagy, P.: Studies of lognormality on mortality statistics. Biometrical Journal (in press)
[14] Kempton, R. A. 1979. The structure of species abundance and measurment of diversity. Biometrics **35**: 307 – 321.
[15] Leiber, B. 1975. Die Nosologie auf dem Wege zu neuen Ordnunssystemen: Syndrome, und Syndromatologie. Internist **16**: 56 – 60.
[16] Pielou, E. C. 1975. Ecological diversity. Wiley, New York – London – Sydney – Toronto.
[17] Preston, F. W. 1948. The commonnest and rarity of species. Ecology **29**: 254 – 283.
[18] de Solla Price, D. 1971. Little Science, Big Science. Columbia Univ. Press, New York – London.
[19] Vital Statistics of the U.S. 1974. Vol. 2, Mortality. U.S. Department of Health, Education and Welfare, Public Health Service, Govt. Print. Off., Washington (1978) pp.
[20] Vital Statistics of the U.S.1975. Vol.2, Mortality. U.S. Department of Health Education and Welfare, Public Health Service, Hyattsville (1979) pp. 1 – 186 – 1 – 251.
[21] Williams, C. B. 1969. Patterns in the Balance of Nature. Academic Press, London and New York.
[22] Zipf, G. K. 1949. The psychobiology of language. Houghton Mifflin, Human Behaviour and the Principle of Least Effort. Addison-Wisley Press.