TRANSACTIONS OF SOCIETY OF ACTUARIES
1950 REPORTS VOL. 2 NO. 4

ON THE LARGE SAMPLE DISTRIBUTION OF MORTALITY RATES BASED ON STATISTICALLY INDEPENDENT LIVES

JOHN E. WALSH

INTRODUCTION

In deriving the large sample distribution of an observed mortality rate based on lives, it is often assumed that

(a) the probability of death within the time interval considered is the same for each person of the investigation,

(b) the individuals of the investigation represent statistically independent observations.

Then it is easily shown that for large n the distribution of

$$(q' - q) \Big/ \sqrt{\frac{q'(1 - q')}{n}} \tag{1}$$

is nearly standard normal (zero mean, unit standard deviation); *i.e.*, for large n the distribution of q' is approximately normal with mean q and standard deviation $\sqrt{q'(1 - q')/n}$. Here n is the number of persons under observation; q' is the observed rate of mortality (number of deaths divided by n); and q is the true value of the rate of mortality (expected value of q').

For practical cases, however, assumption (a) is often of doubtful validity, even in the sense of being a rough approximation to the true situation. This raises the question of how much the large sample distribution of expression (1) depends on (a). This note shows that ordinarily the distribution of (1) is nearly standard normal for large n if only assumption (b) is satisfied; *i.e.*, violation of (a) has little effect on the large sample distribution of q' when (b) holds. By definition, q is the average probability of death for the individuals of the investigation when (a) is not satisfied.

ANALYSIS

The analysis presented is based on several standard but nonelementary theorems of mathematical statistics. For convenience of reference, the theorems used are stated in the Appendix. The derivations for these theorems are not presented in this note. Instead, the reader is referred to these derivations in a textbook on mathematical statistics (*i.e.*, reference [1]).

Let the n persons of the investigation be observed throughout some specified interval of time. These individuals are assumed to represent statistically independent observations for this period of time. Denote the (unknown) probability of death during this time interval for the ith person by q_i, $(i = 1, \ldots, n)$. Then the rate of mortality q is given by

$$q = \sum_1^n q_i/n .$$

It is assumed that the q_i do not tend to either zero or unity as n increases; *i.e.*, that the value of

$$\sum_1^n q_i(1 - q_i)/n$$

does not tend to zero as $n \to \infty$. This seems to be a reasonable assumption for the usual type of practical situation.

First let us consider the asymptotic $(n \to \infty)$ distribution of the statistic q'. The random variable associated with the ith individual, $(i = 1, \ldots, n)$, is denoted by x_i and can only take on the values 0 and 1. The value of x_i is 1 if the ith person dies during the observation period and is 0 otherwise. Actually, x_i is a sample of size 1 from a binomial population for which q_i is the probability of a "success." Thus the mean m_i and variance σ_i^2 of the random variable x_i are given by

$$m_i = q_i , \qquad \sigma_i^2 = q_i(1 - q_i) , \qquad (i = 1, \ldots, n) ,$$

while the third absolute moment of x_i about m_i (*i.e.*, the expected value of $|x_i - m_i|^3$) has the value

$$\rho_i^3 = q_i(1 - q_i)[q_i^2 + (1 - q_i)^2] \leq q_i(1 - q_i) .$$

From this it follows that

$$\Big(\sum_1^n \rho_i^3\Big)^{1/3} \Big/ \Big(\sum_1^n \sigma_i^2\Big)^{1/2} \leq \Big[\sum_1^n q_i(1 - q_i)\Big]^{-1/6} .$$

Hence, since the q_i do not tend to zero or unity as n increases,

$$\lim_{n \to \infty} \Big(\sum_1^n \rho_i^3\Big)^{1/3} \Big/ \Big(\sum_1^n \sigma_i^2\Big)^{1/2} = 0 .$$

This result, combined with the independence of the x_i, shows that the Central Limit Theorem stated in the Appendix is applicable to

$$q' = \sum_1^n x_i/n .$$

As $n \to \infty$, the distribution of ξ_n tends to the standard normal while η_n converges in probability to the constant

$$\sqrt{\frac{\sum_1^n q_i/n - \left(\sum_1^n q_i/n\right)^2}{\sum_1^n q_i/n - \sum_1^n q_i^2/n}}.$$

Using Cramér's Convergence Theorem (see Appendix), the asymptotic distribution of (1) is seen to be normal with zero mean and standard deviation

$$\sqrt{\frac{\sum_1^n q_i/n - \sum_1^n q_i^2/n}{\sum_1^n q_i/n - \left(\sum_1^n q_i/n\right)^2}}. \tag{3}$$

Let us examine the value of (3) for the practically important situations where the q_i are small (say, less than .2). Then $\Sigma q_i/n$ is noticeably greater than either of $\Sigma q_i^2/n$ and $(\Sigma q_i/n)^2$. Moreover, the average of the squares, $\Sigma q_i^2/n$, should have a value somewhere near that of the square of the averages, $(\Sigma q_i/n)^2$. Consequently the value of

$$\frac{\Sigma q_i/n - \Sigma q_i^2/n}{\Sigma q_i/n - (\Sigma q_i/n)^2} \tag{4}$$

should be near unity. Taking the square root of this quantity to obtain (3) brings the value nearer to unity. For example, if the value of (4) lies in the interval .9 to 1.1, the value of (3) lies in the interval .95 to 1.05. Thus, for the usual practical situation, the value of (3) is very near to unity. This verifies the statement in the Introduction.

It should be pointed out that requiring the persons to be observed throughout a specified time interval is not necessary for the results of the Analysis to hold. The large sample distribution of (1) is nearly standard normal under much more general conditions. It is sufficient that the observations be statistically independent and that the q_i are small and do not converge to zero. Each person could be observed during a different interval of time.

APPENDIX

This section contains a statement of the three theorems used in the Analysis. These theorems (in slightly different forms) are contained in reference [1]. Page references to [1] are presented with each theorem.

CENTRAL LIMIT THEOREM (LIAPOUNOFF). *Let $x_1, x_2, \ldots, x_n$ be independent random variables, and denote by m_i and σ_i^2 the mean and variance of x_i, $(i = 1, \ldots, n)$. Suppose that ρ_i^3, the third absolute moment of x_i about its mean (i.e., the expected value of $|x_i - m_i|^3$), is finite for all i. If the condition*

$$\lim_{n \to \infty} \left(\sum_1^n \rho_i^3 \right)^{1/3} \Big/ \left(\sum_1^n \sigma_i^2 \right)^{1/2} = 0$$

is satisfied, then the asymptotic distribution of

$$\frac{\sqrt{n}\left(\sum_1^n x \,/n - \sum_1^n m_i/n \right)}{\sqrt{\sum_1^n \sigma_i^2/n}}$$

is standard normal.

This version of the Central Limit Theorem is presented and verified on pp. 215–17 of [1].

TCHEBYCHEFF'S THEOREM. *Let $y_1, y_2, \ldots$ be random variables, and let m_n and σ_n denote the mean and standard deviation of y_n. If $\sigma_n \to 0$ as $n \to \infty$, then $y_n - m_n$ converges in probability to zero.*

Tchebycheff's Theorem is stated and proved on p. 253 of reference [1].

CRAMÉR'S CONVERGENCE THEOREM. *Let $\xi_1, \xi_2, \ldots$ be a sequence of random variables with the distribution functions $F_1, F_2, \ldots$. Suppose that $F_n(x)$ tends to a distribution function $F(x)$ as $n \to \infty$.*

Let $\eta_1, \eta_2, \ldots$ be another sequence of random variables, and suppose that η_n converges in probability to a positive constant c. Put

$$Z_n = \xi_n / \eta_n .$$

Then the distribution function of Z_n tends to $F(cx)$ as $n \to \infty$.

This convergence theorem is presented and proved on pp. 254–55 of [1].

REFERENCE

[1] Harald Cramér, *Mathematical Methods of Statistics*, Princeton Univ. Press, 1946.

DISCUSSION OF PRECEDING PAPER

ADITYA PRAKASH:

This is an interesting paper. Part of the result obtained by Mr. Walsh can be anticipated by general reasoning. The arithmetic mean of a population, however diverse its elements, is a definite though often indeterminable quantity. The larger a sample gets in size, the smaller is the range of possible differences between the population mean and the sample mean. That this distribution of sample means tends to the "normal distribution" has been frequently observed and commonly assumed by statisticians. Mr. Walsh, however, takes this out of the field of expert guesses to the domain of a mathematically justifiable assumption, in so far as the distribution of sample mortality rates for small $q > 0$ is concerned.

The upper limit of q as .2, mentioned purely as an example in the paper, is sufficiently large to cover most of the mortality table of standard lives and quite a few of the insurable substandard mortality group. Since p is always a complement of q, it follows that if the assumption of a normal distribution is good for $0 < a \leqslant q < a$, then it is also good for $1 - a < p \leqslant 1 - a < 1$. Hence the approximation is good not only for small values of $q > 0$ but also for sufficiently large values of $q < 1$.

Underlying the two assumptions of homogeneity and of statistical independence of the occurrences, specifically mentioned in the first paragraph of the paper, is the assumption of a random sample, and Mr. Walsh, with good justification, takes this for granted in all his derivations. This is mentioned because the usual methods employed for getting a sample of mortality experience from insured lives do not give a random sample when condition of homogeneity is relaxed. If a sample of, say, 100,000 lives is taken for the combined experience of calendar years 1900, 1910, and 1920 and if the number of individuals exposed in each calendar year was greater than 100,000, then it should be possible, in a random process of obtaining a sample, to get an entire sample from the year 1900, or 1910, or 1920, comprising not a single case from the other years. This is not true in the usual procedures[1] of taking as sample the entire experience of a few years or of taking every tenth or twentieth policy. Here our basic data are already arranged by policy numbers, that is, for the most part, by time at issue. If the experience extends over any period during which there have been some basic changes in mortality and not mere chance fluctuations, then as the data are already ordered our methods give us a stratified and not a random sample. Thus we find that, even

[1] In practice, amounts or policies are considered instead of lives. These problems are not considered here, as they are not covered by the paper.

though the variance formulae are derived without assuming homogeneity, in practice we would not be justified in applying these formulae to samples derived according to the methods in common use if no thought is given to the time interval and to consequent changes in mortality experience.

It is quite possible that the variance for stratified sample means may be smaller than the one obtained on assumption of a random sample, but that is a different point.

For the benefit of fellow students it may be recalled that, even if homogeneity be not necessary for a proper estimate of the variance, it is essential for a proper interpretation of the estimated mortality rate and is a valuable safeguard against spurious results. Otherwise there may be some point in modifying our sample designs sufficiently so as to eliminate the aforesaid objections.

Though Mr. Walsh may not have saved the actuarial investigators problems arising from the need for homogeneity, he has broken fresh ground in an interesting direction.

(AUTHOR'S REVIEW OF DISCUSSION)

JOHN E. WALSH:

Mr. Prakash's discussion touches upon several basic points which are only mentioned in the paper. I am grateful for the opportunity of discussing some of these points in more detail.

The problem posed in the paper assumes that certain quantities are given. These include the n persons of the investigation and the time intervals during which these persons are exposed. The procedure which yielded this particular combination of persons and time intervals is not considered; it is not pertinent to the analysis if the conditions specified in the paper are satisfied. As Mr. Prakash points out, however, this procedure can be important in applications. The reason is that the average probability of death for the n individuals may not be the quantity of interest if the individuals and their intervals of exposure are not selected in a suitable fashion. If the results of the paper are to be valuable, it is important to use a procedure for selecting the persons and time intervals which assures that $\Sigma q_i/n$ is the quantity of interest.

It might be worth while to emphasize the generality of the results of the paper. The people considered may be exposed to entirely different conditions. Some may be exposed for longer time intervals than others. They may be of widely varying ages. They may even be born several hundred years apart. If the persons are statistically independent and the probabilities of death are not too large, however, the results of the paper can be used to obtain probability information about $\Sigma q_i/n$.