On Mathematical Statistics and its application to Political Economy and Insurance. By DR. THEODOR WITTSTEIN, Actuary of the Hanover Life Insurance Company. Translated by T. B. SPRAGUE, M.A.

PART I.

INTRODUCTION.

Concerning a New Science, being a Paper read before the Mathematico-Physical Section of the 40th Natural Philosophy Congress, held at Hanover in 1865.

IT is well known that the number of the physical sciences is continually increasing, by a process of development from within. We constantly see divisions and branches of existing sciences

cultivated and developed until they acquire a separate existence as new and independent sciences. For it has long been impossible for a single person to master the whole, or even a considerable part of these sciences, and the necessity for a division of labour is continually becoming more and more felt here, as elsewhere. It is, however, comparatively seldom that the number of the sciences is increased by the addition to them of an entirely new one, and therefore the case of which I am about to speak appears to deserve our particular attention. For I shall presently prove that a department of science, which has not hitherto belonged to the physical sciences, is now being so developed and perfected, that in a very short time it will claim admission into the number on equal terms with the others. This department of science is *Statistics*. The science of Statistics in its present form has now existed a full century. It owed its birth, as well as its name, to Achenwall of Göttingen, about the middle of the last century. It may be defined as the compilation and comparison of all the noteworthy facts that a State—or, in more general terms, a Society—presents to our view at a given epoch; of what Schlözer designates by a single word “*Staatsmerkwürdigkeiten*.” But to this definition the subsidiary idea soon attached, that the information in question must be given principally in numbers; and thus we see that the present text books on statistics really consist, for the most part, of a collection of tables giving all the information that can be expressed in figures, as to the population, manufactures, agriculture, trade, &c., &c., of a country. In this form, statistics is one of the social sciences, and is especially considered as an important auxiliary to the science of political economy. Now we must bear in mind that, philosophically considered, a collection of tables containing observed facts can make no claim to be considered a science in the proper sense of the word. Such tables only furnish the rough materials from which the science is to be constructed. In scientific phrase, they are only a series of *observations*, this very word expressing what is the next step required. The problem before us is the same as in all other branches of natural philosophy—to ascend from observations to the discovery of natural laws. In this direction, statistics, faithful to its definition, has hitherto done scarcely anything. It has, in fact, only drawn the obvious conclusions which result from a first glance at the tables, and, therefore, when it attempts to solve the above problem, it becomes essentially a new science. It has to begin at the point where statistics, as hitherto understood, ceases; and since the new science deals principally with figures, we must

employ mathematics for the solution of the problem. The new science may therefore be called *Mathematical Statistics*, unless mathematicians should prefer to call it *Analytical Statistics*, according to the analogy of Analytical Optics, Analytical Mechanics, &c.

This will be made still clearer by a comparison with another instance of which the circumstances are more generally known. The astronomer makes observations and collects them in the form of tables; but if he regarded his work as then completed, astronomy could never have claimed to be called a science. It would have remained simply in the stage occupied by statistics as hitherto understood. Such nearly was the position of astronomy when Tycho Brahe made his observations on the planet Mars, which have since become so celebrated. Then came Kepler, who from these observations deduced his well known laws. In Kästner's happy phrase, "Tycho quarried the marble, and Kepler chiseled the statue." It was this step, afterwards completed by Newton's theoretical investigations, that first raised astronomy to the rank of a science, properly so called; and this step has not yet been made by statistics. It must not, however, be supposed that a Kepler only is required to raise statistics to the dignity of a science. Its Tycho has not yet appeared; for statistical facts and figures are as yet, with scarcely an exception, of very doubtful value, and (as I shall presently show) unsuited for mathematical investigation. Nay, keeping to the same metaphor, it may be said that its Copernicus is still to come, who shall for the first time give us a general view of the facts we ought to observe. Statistics is now in its infancy—at the same stage as astronomy was when it was nothing but astrology and drew horoscopes. It still happens every day, that people prove by statistical figures whatever they choose, the only thing necessary for the purpose being a kind of dexterity, happily expressed in the French phrase, "*grouper les nombres*."* This will be amended as soon as mathematics, with its inexorable logic, shall have mastered the statistical materials. The calculus of probabilities is the special branch that must here be applied, and that not in the sense so commonly attached to the phrase by statisticians, according to which it is a calculus that gives *inexact* results as distinguished from exact calculations, but that ingenious science which is as completely based on logic as any part of

* Say gives one of many instances in his *Traité d'économie politique*: "The French Minister of the Interior, in his report for 1813, a time of calamity, when commerce was destroyed and the national resources of every kind were rapidly diminishing, boasts of having demonstrated by means of figures that France was in a condition of prosperity greater than it had ever before enjoyed."

mathematics, and which has been brought to so splendid a development by Laplace and Gauss. Only thus will statistics grow to be a science, in the full meaning of the word; and only thus will really useful and trustworthy results be obtained. It may even be predicted that, in a future century, Mathematical Statistics will solve problems of the bare enunciation of which we have as yet no idea.

What has been here said of statistics in general, is not equally true of all the divisions of the science; for it is only the statistics of population which now manifestly admits of mathematical treatment, whereas the other divisions must for the present be excluded from it. In the statistics of population some steps have indeed been already made, which may be considered as an attempt to lay the foundation of mathematical statistics, but these steps are so far from satisfactory, and have made so little use of the means of investigation which analysis in its present advanced condition could have supplied, that it is necessary to disregard them and commence the investigation anew from the very beginning.

The first and most important idea with which the mathematical treatment of the statistics of a population has to deal, is that of mortality; for this enters into every question that can be proposed with regard to a population. For the purpose of answering enquiries as to the mortality that will prevail among a given population or any other given body of persons, it has been usual to form what is known as a mortality table, which represents the rate at which that body of persons will die off, on the supposition that the rate of mortality prevailing at one instant of time will remain unchanged thereafter. A large number of such mortality tables are to be found collected in statistical text books, but they are all, more or less, of doubtful value. Various attempts have also been made to find a formula that shall represent the number living at any age in the table; but these attempts were certain to fail, so long as the fundamental facts were deficient in the accuracy which, as stated above, was characteristic of Tycho's observations. Even the function, first proposed by Gompertz and since discussed by so many persons, $y = a^{b^x}$, which leads in its application to logarithmic integrals and gamma functions, must remain a very doubtful hypothesis until a sufficiently trustworthy collection of facts can be obtained to test its accuracy.

It is, indeed, no exaggeration to assert that all existing mortality tables, without exception, are far from being so trustworthy as could be desired. This is owing, partly to the imperfection of

the facts on which they are based, partly to the imperfection of the methods hitherto employed in their construction. As regards the former, the censuses in Germany are at present taken in so unsatisfactory a manner, that calculations based on them can only be considered as rough approximations to the truth. This is, at all events, true as regards the censuses within the limits of the Zollverein, whereas in Belgium and France, where to be sure they go to greater expense, matters appear to be in a better condition. Having elsewhere stated my views as to what science requires in a census, I will not repeat them here,* but will only add that, unfortunately, the public are but little disposed to assist in the matter. It is, for instance, well known that women universally state their ages too low; nay, at the last census in Hanover, two ladies positively refused to state their ages at all, so that the enumerators could do nothing but guess at them. Under such circumstances, we must, perhaps, give up all idea of ever obtaining a completely trustworthy census. More accurate data can be obtained from the experience of Life Insurance Companies, Annuity Companies, Widows' Funds, and similar societies, but hitherto very few of these have been at the pains to extract and arrange, with a view to strict mathematical treatment, the statistical facts lying buried in their records; and we therefore possess much fewer mortality tables derived from such sources than might have been expected. But to tables of this kind the second of the above mentioned objections applies; for they are defective, in consequence of the defective methods employed in their construction. Even the two which at the present time are considered the best, and are extensively used, namely, Brune's Table, deduced from the experience of the General Widows' Fund, in Berlin, and that deduced from the experience of the 17 English Offices, are open to this objection; for the methods employed in their construction, so far as they are known, are far from being theoretically correct.

It must be admitted that the labour bestowed by statisticians on the formation of mortality tables is so far justified, that a mortality table presents in a very clear and intelligible form the facts to which it relates, and exhibits extremely well the instantaneous changes in a population. But, after all, it is only a popular form, insufficient for scientific purposes, and it has accordingly been forcibly pointed out by other writers† that the formation of a

* See my Essay, *Zur Bevölkerungs-Statistik*, in the *Zeitschrift des Königlich Preussischen statistischen Bureau*, 3rd year, Part I.

† See, especially, Fischer's "*Grundzüge des auf die menschliche Sterblichkeit gegründeten Versicherungswesens*," Oppenheim, 1860.

mortality table cannot be the first step of scientific enquiry. The fundamental idea from which our treatment of the statistics of a population must start, is rather that of the *probability of dying*, or more exactly, *the probability that a person who belongs to a definite group will die within a year*. Statistics has hitherto lacked this idea, because, as already remarked, it has not had a correct idea of the calculus of probabilities in general; and with the introduction of the above idea, the science becomes essentially a new one. We can, if we please, substitute for the above probability, that of living a year, which is the complement to unity of it. If either the one or the other of these probabilities is known for every year of age, it requires but little labour to deduce a mortality table.

The analytical expedients required for the solution of the problem have been so fully developed by Laplace, in his *Théorie Analytique des Probabilités*, that it is rather surprising they have not been earlier applied to it. It must be confessed, however, that it is but lately that the want has become fully felt, in consequence of various recent writers having attempted the solution of the questions under consideration. These have certainly, one after the other, made steps towards it, but have not yet reached the desired goal. We shall here endeavour, with the help of the above mentioned expedients, to make another step forward.

If it has been observed that out of a group of L persons, there are L' alive at the end of a year, and there have been no entries into this group, nor withdrawals from it during the year (in which cases a special investigation would be required), it has been the custom hitherto to say that the probability, p , of living a year is given for each person of this group by the equation

$$p = \frac{L'}{L}.$$

But it is not forgotten to add that if, under the like circumstances, other values of L and L' were given, the value of p may generally be different. Hence it logically follows that the calculation gives the value of p only inexactly; and thus originates the strange idea prevailing among the unlearned that the calculus of probabilities gives inexact results. Now, in reality, the above form of expression is wholly incorrect. The value of p given by the above equation makes no claim whatever to be the *true value* of the unknown probability, but it is the *most probable value* resulting from the observations that have been made. Now to the most probable value there always belongs a *probable error*, and a *mean*

error, either of which may be used to measure the degree of trustworthiness of that most probable value. This probable and this mean error have, however, never yet been determined.

Again, if the probability p of living a year is given, and we wish, under the same circumstances as before, to know the number, l' , which will be alive at the end of a year out of a new group of l persons, it has hitherto been the custom to say that it is given by the equation,

$$l' = lp.$$

At the same time it is admitted, that actual observation may show a number different from l' , so that in this case too the calculation may be charged with being inexact. But here, also, the usual form of expression, as given above, is inaccurate; for the value of l' given by the equation is not the *true value*, but the *most probable value* of the number alive at the end of a year. Hence this value also has a *probable error*, and a *mean error*, either of which measures its degree of trustworthiness; and these probable and mean errors have likewise never yet been determined.

These two cases furnish examples of a few of the gaps which statistics, as hitherto understood, has in vain sought to fill up. To fill them up, and to build further on the ground thus gained, is the object of the following investigations, which will be found to take an entirely new direction. The degree of precision in our treatment of the subject, and the important applications it admits of, will be seen so directly from the investigations themselves, that it appears unnecessary to say more of them here.

PART II.

GENERAL INVESTIGATIONS AS TO MORTALITY AND MORTALITY TABLES.

Chapter 1.

The probability of living a year is supposed to be given.

§ 1.

Let p be the probability that a person of the age x will be alive at the end of a year.* This is most commonly given by means of

* It may be remarked, in passing, that the following investigations admit of being applied to many other subjects besides mortality. Thus, for instance, we might have said, "Let p be the probability that an unmarried person of the age x will still be unmarried at the end of a year;" or "Let p be the probability that a building of class x will not be burnt down at the end of a year," and so on; the nature of the particular subject indicating what modifications must be introduced in applying our investigations to such cases as these. We will not pursue this subject further on the present occasion.

a mortality table; the ratios of the numbers living at successive ages being equal to the different values of this probability. Then, according to well known principles, the probability that out of l persons of the age x , l' will be alive at the end of a year, is

$$P = \frac{l}{l'} p^{l'} (1-p)^{l-l'} \quad . \quad . \quad . \quad (1)$$

Here l' can have any integral value from 0 to l inclusive.

The value of l' which makes P a maximum can be found by comparing the above value of P with the values it receives when $l'-1$ and $l'+1$ are substituted for l' . It is given by the relations,

$$\frac{l-l'+1}{l'} > \frac{1-p}{p} > \frac{l-l'}{l'+1};$$

$$i. e. \quad \frac{l'+1}{l+1} > p > \frac{l'}{l+1};$$

whence, if l is a large number, and l_0 denotes the most probable value of l' , we have

$$l_0 = lp. \quad . \quad . \quad . \quad . \quad . \quad (2)$$

This is the number given in a mortality table as living at the age $x+1$, if l is the number living at the age x .

§ 2.

Putting $l' = l_0 + k = lp + k$, then by (1) the probability that the number alive at the end of a year will exceed by k the most probable number, is

$$P = \frac{l}{lp+k} \cdot p^{lp+k} \cdot (1-p)^{l(1-p)-k} \quad . \quad . \quad (3)$$

Here k may have any integral value from $-lp$ to $l(1-p)$.

If we assume that l , $lp+k$, $l(1-p)-k$ are large numbers, and apply Stirling's formula

$$n! = \sqrt{2\pi} \cdot n^{n+\frac{1}{2}} \cdot e^{-n}, \text{ approximately,}$$

equation (3) becomes

$$\begin{aligned} P &= \frac{1}{\sqrt{2\pi}} \cdot \frac{l^{\frac{l}{2}}}{(lp+k)^{lp+k+\frac{1}{2}} \{l(1-p)-k\}^{l(1-p)-k+\frac{1}{2}}} \cdot p^{lp+k} \cdot (1-p)^{l(1-p)-k} \\ &= \frac{1}{\sqrt{2lp(1-p)\pi}} \cdot \frac{1}{\sqrt{\left(1+\frac{k}{lp}\right) \left\{1-\frac{k}{l(1-p)}\right\}}} \cdot \\ &\quad \frac{1}{\left(1+\frac{k}{lp}\right)^{lp+k} \left\{1-\frac{k}{l(1-p)}\right\}^{l(1-p)-k}}. \end{aligned}$$

Expanding the second factor by the Binomial Theorem, it becomes

$$1 + \frac{(2p-1)k}{2lp(1-p)} + \dots$$

and the Napierian logarithm of the third factor is

$$-(lp+k) \left(\frac{k}{lp} - \frac{1}{2} \frac{k^2}{l^2 p^2} + \dots \right) \\ - \{l(1-p)-k\} \left\{ -\frac{k}{l(1-p)} - \frac{1}{2} \frac{k^2}{l^2 (1-p)^2} - \dots \right\}$$

or

$$-\frac{k^2}{2lp(1-p)} + \dots$$

the other terms containing powers of l in the denominator, and therefore vanishing in comparison with the first.

If, then, for brevity, we put

$$\frac{1}{2lp(1-p)} = h^2 \quad . \quad . \quad . \quad . \quad . \quad (4)$$

the above formula becomes

$$P = \frac{h}{\sqrt{\pi}} \{1 + (2p-1)h^2 k\} e^{-h^2 k^2} \quad . \quad . \quad . \quad (5)$$

or finally, since we consider k to be small in comparison with l ,

$$P = \frac{h}{\sqrt{\pi}} \cdot e^{-h^2 k^2} \quad . \quad . \quad . \quad . \quad . \quad (6)$$

It may be noticed, that it hence follows, since the extreme values of which k is capable extend from $-\infty$ to $+\infty$, that

$$\int_{-\infty}^{\infty} P dk = 1.$$

Observation. The following numerical examples will enable the reader to judge of the accuracy of Stirling's formula applied above.

n	Log $\lfloor n$		Error.
	By actual Calculation.	By the Formula.	
5	2.07918	2.07195	+ .00723
10	6.55976	6.55615	.00361
25	25.19065	25.18920	.00145
50	64.48307	64.48235	.00072
100	157.97000	157.96964	.00036

From this it is seen that even for values of n that are not large, the error of the five-figure logarithm affects only such part

of it as we should consider of doubtful accuracy on account of the uncertainty of our data. But it would be a wholly needless attempt at exactness to use more than 5 places of decimals in calculations of this kind.

§ 3.

The probability that the number living at the end of a year will lie between $lp+k$ and $lp-k$ is by (6)

$$= \frac{2h}{\sqrt{\pi}} \int_0^k e^{-h^2 k^2} dk \quad . \quad . \quad . \quad (7)$$

If this probability is to be equal to $\frac{1}{2}$, and σ is the value of k for which this happens, we have, as is well known,*

$$\sigma = \frac{\cdot 6745}{h\sqrt{2}};$$

whence, by means of (4),

$$\sigma = \cdot 6745 \sqrt{lp(1-p)} \quad . \quad . \quad . \quad (8)$$

and this is the *probable error* to be expected in the observed number l' ; or, in other words, it is an even bet that the number living at the end of a year will lie between the limits

$$lp \pm \cdot 6745 \sqrt{lp(1-p)}.$$

The quantity $\frac{1}{h\sqrt{2}}$, which is here $= \sqrt{lp(1-p)}$, represents, as is well known, the *mean error* of the observations.

It follows that, the mortality being the same, the probable error in the number living at the end of a year is proportional to the square root of the number observed. On the contrary, the ratio $\frac{\sigma}{l}$, that is, the ratio that the probable error bears to the number observed, is inversely proportional to the square root of the number observed, and can therefore be made as small as we choose, by taking that number sufficiently large. These conclusions hold also for the mean error, and for every other characteristic error. The probable error in the number living at the end of a year is evidently identical with the probable error in the number dying within the year.

§ 4.

If l_0 , the most probable number living at the end of a year, is known beforehand, as is the case in calculations based upon a given

* See Gauss's *Theoria combinationis observationum*. The coefficient $\cdot 6745$ is more exactly $\cdot 6744897$, and its five-figure logarithm is $\bar{1} \cdot 82898$.

mortality table (see the end of § 1), the equation (8) can be put in the form

$$\sigma = .6745 \sqrt{\frac{l_0(l-l_0)}{l}} \quad . \quad . \quad . \quad . \quad . \quad . \quad (9)$$

that is, it is an even chance that the number living at the end of a year will lie between the limits

$$l_0 \pm .6745 \sqrt{\frac{l_0(l-l_0)}{l}}.$$

For instance, according to Brune's table, out of 7943 males of the age 40, 7847 are alive at the end of a year. If we put these numbers respectively for l and l_0 in the above formula, we get $\sigma = 6.57$; and it is an even chance that if 7943 men of the age 40 are under observation, the number living at the end of a year will lie between 7840.43 and 7853.57.

Observation. If the probability (7) is to have the values .9, .99, .999, .9999, the coefficient .6745 in (8) and (9) is replaced by 1.6449, 2.5758, 3.2918, 3.8906, respectively. (See Gauss, *Theoria comb. obs.*) Consequently the odds are 9 to 1 that the number living at the end of a year will lie between the limits

$$lp \pm 1.6449 \sqrt{lp(1-p)},$$

or

$$l_0 \pm 1.6449 \sqrt{\frac{l_0(l-l_0)}{l}}.$$

In other words, it is to be expected in a large number of observations that, in every 10 cases in which l such persons are observed, the number living at the end of a year will only once fall outside these limits. Similarly in the other cases.

§ 5.

The foregoing calculation is not confined to a single year, but can be at once applied to any term of years for which the value of p is given. In particular, starting from the number living at any age we choose, in a given table of mortality, we can immediately calculate the probable error in the number living at any higher age, by substituting for l_0 in formula (9) the number living at that age according to the table.

For instance, taking Brune's table for males, in which 7943 is the number living at the age 40, we get the following probable errors in the number living at various ages.

Age.	Number Living.	PROBABLE ERROR IN THE NUMBER LIVING.	
		Absolute.	Percentage of Number Living
40	7943		
50	6845	21	·3
60	5304	28	·5
70	3100	29	·9
80	954	20	2·1
90	46	5	10·8

The probable error in the number living attains its maximum value when $l_0 = \frac{l}{2}$, that is, for the probable duration of life (in the above table for the age of 66, for which $\sigma = 30$); and beyond that point diminishes down to 0. On the contrary, the ratio $\frac{\sigma}{l_0}$, or the ratio that the probable error in the number living bears to that number, continually increases.

(To be continued.)