ON THE STATISTICAL MEASURE OF INFECTIOUSNESS.

By M. GREENWOOD, F.R.S.

We all recognise that some diseases are more "catching" than others. Every mother knows that measles is very catching and most people set aside a group of common complaints, measles, mumps, whooping-cough, scarlet fever, diphtheria—perhaps roughly in that order—as catching complaints. Then again, still keeping ourselves within the circle of ideas of educated non-medical people, one has such complaints as common colds or influenza which one thinks of as running through a house indeed but does not put quite into the measles category, as one feels that factors determine the spread other than mere proximity to a sick person. Lastly, one has some illnesses, gonorrhoea would be a fair example, which everybody recognises to be spread wholly by contagion, almost always by a particular method of contagion, but does not regard as catching at all in the sense that measles and whooping-cough are catching. When we enquire into the reasons of these opinions they will be found, I think, to be these.

An illness is held to be catching when it has usually been possible to explain the existence of a case of it by close association (of some kind) with an immediately pre-existing case; the notion of more or less infectiousness depends upon some appraisement of the proportion of persons attacked to persons exposed to the risk of attack. Mothers observe that when a child sickens with measles most of the other children take the complaint within a period of days or weeks, when a child sickens with scarlet fever the proportion of others attacked is smaller, and so on.

When we are dealing practically with such complaints as measles, rough criteria such as these are sufficient; but when we seek a deeper knowledge of the epidemiology even of these common complaints they are inadequate, and when we pass to the debatable region of diseases of which the infectiousness is quite uncertain they are altogether useless. Epidemiologists without any predilections for statistical analysis have, therefore, been forced to give these popular notions an at least quasi-statistical form. They have studied the distribution of "multiple" cases of disease in groups—families, inmates of houses, etc.—and sought to draw conclusions from the frequency distribution. If, they have argued, a "case" does not breed other "cases" in the exposed to risk, then if a number of groups have been observed through a finite interval, we shall find but few instances in which any one group contains more than one affected member, and in those few cases we shall be able to account for the independent origin—independent one of another—of the multiple cases. If, on the other hand, one case does breed others, the more intense the effect the

more closely will the distribution of groups tend to approximate to the opposite extreme, viz. n cases in a group of n persons. The criterion of infectiousness is the degree of approximation to one or other extreme J-shaped distribution, only a single case in each group of n, n cases in each such group. Thus the fact that in 8635 families each containing at least one member with poliomyelitis, 96 per cent. had *only* one affected member, has been used as an argument against the belief that personal contact is an important factor of dissemination of this complaint[1].

The object of this paper is to examine the statistical problem in greater detail. There are two difficulties, one really insuperable, the other insuperable by me.

The generally insuperable difficulty (insuperable in the present state of knowledge) is that *accurate* data are scanty; a great majority of the published records of multiple cases omit essential particulars. It is evidently futile to apply elaborate arithmetical machinery to records of "multiple" cases which omit any appraisement either of the characteristics (age, sex, previous history of exposure) or even of the numbers constituting the groups. In the rare instances when this information is afforded, the number of groups is tiny. Had it not been for the kindness of my friend Dr Stocks, this paper must have been wholly "in the air." The difficulty insuperable by me is that before one has given much time to the subject—I have worked at it, off and on, for years——one encounters mathematical difficulties of a serious nature, the conquest of which requires special knowledge and ability I do not possess.

These considerations are by way of preface to and apology for the report of an investigation practically and theoretically inadequate but, as I hope, suggestive.

The general problem is this. N households (families or groups defined in some precise way) have been observed over a period of time T and, at the close of observations, N_0 of the groups have recorded no case of the disease, N_1 have recorded 1 case, N_2 2 cases, ... N_r r cases; what light does the form of this distribution throw upon the aetiology of the cases? The problems to be discussed fall under two classes. In both the number of members of each group is supposed to be constant, but in one the number of cases within a group may, in the other it may not exceed the number of persons constituting the group. To the former class belong such problems as that of "cancer houses" where the number of cases may exceed the number of inhabitants at a given instant, since the record may cover some generations of inhabitants, or of the distribution of non-fatal accidents amongst employees, since one employee may sustain many accidents; to the latter the distribution of cases of measles, scarlet fever, etc., in families or houses when the period of observation T is short. The distinction is clear; in the two former cases, although not more than all the inhabitants can be sick of cancer at one and the same time, and a workman cannot have an unlimited number of accidents, however trivial, at precisely

[1] *Epidem. Report, Health Section, League of Nations*, March 15th, 1930, pp. 100–1.

the same instant of time, yet in the one case the period of observation T may be so long that successive generations of inmates have been exposed to risk, and in the other the period of actual happening is so short in comparison with T that there is no real impropriety in postulating an upper limit to the possible incidence of multiple cases far greater than the actual number of persons. But, when, as always happens in the testing of a hypothesis and sometimes happens in the framing of it, the observed number of *cases*, which I shall always denote by n, is taken to be, like N, an ultimate datum of observation, we must remember that the upper limit is really fixed. The scientific discussion of problems of the first class was begun in 1911–12 by Maynard and Troup and by Pearson. Pearson, discussing cancer houses, used the following argument. Supposing the distribution of cases in houses to be absolutely random, analogous to throwing balls at random into equally accessible pigeon holes, then the chance that any one case will fall in any one house is $1/N$ and that it will not is $(N-1)/N$, therefore the distribution of the whole n cases should be given by the terms of the binomial $N[1/N + (N-1)/N]^n$. In comparing the observed distribution with that required by the hypothesis, we must not use the ordinary Goodness of Fit test, because not only N but also n is fixed; Pearson showed that in our comparison we must not include the frequency of houses with no cases. In 1920 Yule and I discussed the cognate problem of multiple accidents. We pointed out a *theoretical* objection to Pearson's treatment, viz. that when N and n were not large two sets of data, one recording n cases in N houses, the other kn cases in kN houses, would have different representative binomials which seemed objectionable. We suggested that a better analogy than that of throwing n balls in N pigeon holes might be to liken the exposed to risk to targets subjected to bombardment, the time of bombardment T to be divided into intervals so short that within each not more than one hit could be registered on a target. On that hypothesis the distribution of 0, 1, 2, etc., hits at the end of T should be given by the Poisson limit to the binomial with parameter n/N. The hypotheses are slightly different and, as I still think, the second is the more appropriate, but in practice N would usually be large and n/N always finite, so that there would be little arithmetical difference between the results. Both the pigeon-hole schema and the variant Yule and I favoured can be generalised (see Greenwood and Yule (1920), pp. 259, etc.), although only in the simplest case did we reach a form suitable for the deduction of the requisite parameters from the moments of the statistics. The generalisation of the pigeon-hole schema is of some interest, because it suggests a way of approach which, possibly, might repay further mathematical treatment.

When there is no bias, *i.e.* when the fact that a particular pigeon hole has already received one or more of the first r balls neither increases nor diminishes the probability that one or more of the remaining $n-r$ balls shall fall into it, we can deduce the distribution, viz. $N\left(\frac{1}{N} + \frac{N-1}{N}\right)^n$, term by term, as follows. Suppose we use the symbol ${}_rf_s$ to denote the number of pigeon holes

which after r balls have been distributed contain s apiece. For all values of r, $\sum_{s=0}^{s=n} {}_rf_s = N$.

Let E_rf_s denote the mathematical expectation of ${}_rf_s$, *i.e.* the value to which the mean or average value of ${}_rf_s$ will tend when deduced from a large number of observations or experiments. Then E_rf_s will be a constant dependent on N, n, r and s, and hence $E(E_rf_s)$, which we may write $E^2{}_rf_s$, $= E_rf_s$.

When ${}_{r-1}f_0$ is known, ${}_rf_0$ will differ from ${}_{r-1}f_0$ by -1 or 0 according as the rth ball falls into an unoccupied pigeon hole or not. The probability of the first event is ${}_{r-1}f_0/N$ and of the second $1 - \frac{{}_{r-1}f_0}{N}$. Hence the *mathematical expectation* of the difference

$$ {}_rf_s - {}_{r-1}f_s = (-1)\frac{{}_{r-1}f_0}{N} + (0)\left(1 - \frac{{}_{r-1}f_0}{N}\right), $$

or
$$ E({}_rf_0 - {}_{r-1}f_0) = -\frac{1}{N}{}_{r-1}f_0. $$

Hence
$$ E^2{}_rf_0 = E^2{}_{r-1}f_0 - \frac{1}{N}E_{r-1}f_0, $$

or
$$ E_rf_0 = \left(1 - \frac{1}{N}\right)E_{r-1}f_0. $$

But
$$ {}_0f_0 = N, \quad \therefore E_0f_0 = N. $$

Hence
$$ E_nf_0 = \left(1 - \frac{1}{N}\right)^n N = N\left(\frac{N-1}{N}\right)^n. $$

More generally, the rth ball can fall into one of the ${}_{r-1}f_s$ pigeon holes which already contain s balls each. This gives

$$ {}_rf_s - {}_{r-1}f_s = -1 \qquad \text{.....(i)}, $$

or into one of the ${}_{r-1}f_{s-1}$ pigeon holes which contain $s-1$ balls each. This gives

$$ {}_rf_s - {}_{r-1}f_s = +1 \qquad \text{.....(ii)}, $$

or into one of the remaining $N - {}_{r-1}f_s - {}_{r-1}f_{s-1}$ pigeon holes. This gives

$$ {}_rf_s - {}_{r-1}f_s = 0 \qquad \text{.....(iii)}. $$

The probabilities of the three events are

$$ {}_{r-1}f_s/N, \quad {}_{r-1}f_{s-1}/N, \quad 1 - \frac{{}_{r-1}f_s + {}_{r-1}f_{s-1}}{N}. $$

Hence
$$ E({}_rf_s - {}_{r-1}f_s) = -\frac{{}_{r-1}f_s}{N} + \frac{{}_{r-1}f_{s-1}}{N}, $$

and noting again that
$$ E^2{}_rf_s = E_rf_s, $$

we have
$$ E_rf_s = \left(1 - \frac{1}{N}\right)E_{r-1}f_s + \frac{1}{N}E_{r-1}f_{s-1}. $$

The solution of this difference equation for E_rf_s is

$$ E_nf_s = N\binom{n}{s}\frac{1}{N^s}\left(\frac{N-1}{N}\right)^{n-s}. $$

Let us suppose now that the chance of the rth ball falling into an empty pigeon hole differs from $\frac{r-1f_0}{N}$, so that the chance of falling into an occupied pigeon hole is not $1 - \frac{r-1f_0}{N}$ but equal say to $t\left(1 - \frac{r-1f_0}{N}\right)$, so that the chance of falling into an empty pigeon hole becomes

$$1 - t\left(1 - \frac{r-1f_0}{N}\right) = 1 - t + t\frac{r-1f_0}{N};$$

we find in that event

$$E_n f_0 = \frac{N}{t}\left(\frac{N-t}{N}\right)^n + \frac{N}{t}(t-1)$$

and the expectations of the other frequencies may be similarly deduced.

Turning to the bombardment analogy, the generalisation of the Poisson schema for a single bias is almost as simple as that of the pigeon-hole schema and its formal generalisation to 2, 3, ... r biasses very much easier. But, as in the pigeon-hole schema, it only proved possible (for Yule and me; more dexterous or learned mathematicians might succeed) to reach a solution *in terms of moments* for the case of a single bias.

Generalising in a different way, *i.e.* by seeking the form of the distribution when *a priori* the chances of distribution are not equal, one finds the bombardment schema leads to a very elegant solution. If we suppose the individuals (in an accident distribution) or groups to present *a priori* differences of accessibility to bombardment and that this *a priori* distribution is continuous and of skew binomial type (Pearson's Type III) the solution is eminently practicable for statistical work (see Greenwood and Yule, 1920, p. 273).

The corresponding generalisation of the pigeon-hole schema would be much less suitable. We should have to divide the N pigeon holes into groups, determine the probability that the n balls would be partitioned among the groups in a particular way by a multinomial, then work out the distribution within each group of pigeon holes receiving 0, 1, etc., balls. Only if the number of groups were very small would the operation be a practicable one.

On the whole, I think the methods set out in Greenwood and Yule's paper are fairly satisfactory. Thanks to the researches of Ethel M. Newbold, they have been improved into useful tools for the investigation of *accident* statistics. The practical handling of data relating to multiple cases of disease fulfilling the condition is more difficult because it rarely happens that the specification of the data is complete. Even when we are concerned with the statistics of trivial accidents it may be objected that our specification of the N_0 frequency is incomplete or incorrect, that we may have included workpeople who were never at risk at all or have excluded others that were at risk. When we are informed that in N_1 houses one case of cancer was recorded, in N_2 two, and so on, we shall be in doubt as to N_0. If we take for that the whole number of houses in the administrative area less the sum of those in which cases occurred, we shall surely introduce a heterogeneity. If, on the other hand, we decide

to deduce our Poisson constant or our binomial constants from the truncated frequency we are using a method of low efficiency; in fact a layman might urge that we are playing *Hamlet* without the Prince of Denmark for the unaffected groups will usually much outnumber the sum of those affected. It is perhaps worth noting, although the point is not one of very great importance, that if we happen to have truncated data, *i.e.* data from which the N_0 frequency is wanting for groups of different sizes, some light may be thrown upon the applicability of a *binomial* by the following consideration.

Suppose we were dealing with a character distributed with constant chance $p + q = 1$, through groups of size m and km respectively. The proportions of marked persons would naturally be identical (within the limits of error of sampling), viz. p in each series. But the ratios of marked persons to the number of members of groups of which at least one member of each was marked would not be identical. They would in fact be respectively $p/(1 - q^m)$ and $p/(1 - q^{km})$, so that the ratio would diminish as k increased. In other words the *attack rate* would decrease as the size of the group increased. For instance, suppose that some disease were really not infectious at all but fell at random upon 10 per cent. of a population and that statistics were compiled of families of one, two and five children, all the families having at least one affected member. The attack rate in families of one would be 100 per cent., in families of two, 52·6 per cent. and in families of five, 24·4 per cent. The point is a simple one; some arguments respecting the influence of overcrowding upon mortality and morbidity have been weakened by failing to notice it[1].

With these remarks I leave the kind of problem of which multiple non-fatal accidents and "cancer houses" are typical; until a more expert mathematician provides a better solution of the generalised Poisson series than Yule and I obtained, the methods discussed by us 10 years ago seem to me the most effective available.

I pass now to the other class of problem; one has groups of m individuals each and the number of cases cannot exceed m in a group. The data of observation are restricted to groups which contain at least one marked individual. This is much the most important type of problem for the epidemiologist. It has been said that the solution of every statistical problem depends on the preliminary discovery of the appropriate *Urnenschema*. "Professors of probability," said Mr Keynes, "have been often and justly derided for arguing as if nature were an urn containing black and white balls in fixed proportions. Quetelet once declared in so many words: 'l'urne que nous interrogeons, c'est la nature.' But again in the history of science the methods of astrology may prove useful to the astronomer; and it may turn out to be true—reversing Quetelet's expression—that 'la nature que nous interrogeons, c'est une urne[2].'" I think that this deductive method, viz. to imagine a way of happening, to express that mode quantitatively and to compare its consequences with the

[1] *Proc. Roy. Soc. Med.* 1925. Sect. of Epidem. and State Medicine, p. 38.

[2] Keynes, *A Treatise on Probability*, p. 428.

observed arithmetical facts, is the right course here. Current English statistical methods are perhaps *too* inductive. One sometimes spends much labour in "fitting" frequency formulae to observations having the, rather optimistic, *arrière-pensée* that in this way one can reach not merely a neat representation of the facts but also some insight into the mechanism which brought those facts about. Some years ago an English statistician took a German statistician to task because the latter had drawn some conclusions from the apparent appropriateness of Poisson's limiting binomial for the description of certain data. The critic remarked that as good, or better, fits could often be obtained by using a binomial with a fractional or negative exponent and suggested that the interpretation of such binomials needed consideration. The criticised one, in a very heated rejoinder, poured scorn upon the "formalistic" statisticians who trifled with such fantastic notions. Without seeking to meddle in that quarrel, which was a very pretty quarrel as it stood, I prefer only to use formulae here which have, or which at least I think have, a biological interpretation perfectly intelligible. To make the discussion quite clear I base it upon arithmetical examples. My friend Dr Stocks kindly provided me with Table I and we are to try to interpret the frequency distributions it contains.

Table I. *History of children aged under* 10 *years exposed to a case of measles in same house during* 1926 *epidemic in St Pancras.* (*Period at risk from* 4 *days after appearance of rash in first case, until an interval of* 1 *month has elapsed from onset of any case without a fresh case developing**.)

No. attacked during period at risk	No. of contacts under 10 years of age (first case not included)										
	1	2	3	4	5	6	7	8	9	10	Total
0	340	197	84	60	25	11	3	2	1	—	723
1	164	104	60	29	15	6	4	2	—	—	384
2	—	57	57	25	9	4	—	3	—	1	156
3	—	—	27	11	10	3	3	3	—	—	57
4	—	—	—	7	1	—	2	—	—	—	10
5	—	—	—	—	1	1	3	—	—	—	5
6	—	—	—	—	—	1	—	—	—	—	1
7	—	—	—	—	—	—	1	1	—	—	2
Total	504	358	228	132	61	26	16	11	1	1	1338

* A case developing within 3 days of first case was regarded as a simultaneous infection and not included as a contact. Contacts developing measles more than a month from onset of a preceding case were regarded as having escaped attack for purposes of this table.

Let us begin with the simplest (and, of course, certainly erroneous) hypothesis that measles is not an infectious disease at all, but that by picking out the houses in which at least one case occurred we have secured groups which were indeed exposed to some special risk. This might easily happen. A nefarious typhoid carrier might have contaminated the sealed family milk bottles of a particular group of families. Typhoid would be restricted to those families, it *might* attack at least one member of each such family and, if they were all persons who sedulously followed the spirit and letter of hygienic ritual

in eating and drinking, none of the subsequent cases would be due to intramural infection, the distribution must be wholly determined by personal susceptibility and dosage. In that case what sort of a distribution should we expect? Our old friend the simple binomial attracts us. Let us take for its exponent the number in house and for its p the ratio of attacked to exposed to risk. Applying this as an example to the group $m = 3$ we have the computed distribution, which is the third column of Table II. Clearly it does not agree

Table II.

Secondary Cases	Frequency	First binomial	Second binomial	Poisson	Chain binomial
0	84	76·2	56·1	74·6	89·7
1	60	100·9	100·3	83·3	52·7
2	57	44·4	59·7	46·6	54·4
3	27	6·5	11·8	23·5	31·2

at all with observation; there are far too many "multiple" cases observed for the hypothesis. Not even if we base our binomial upon the individuals of groups of $m = 3$ instead of upon the whole experience of all the groups do we reach a tolerable agreement. Certainly that hypothesis may be excluded. As a mere matter of empiricism the methods of the Greenwood and Yule paper were tested upon these data. The results were better than with the binomials, for instance an uncomplicated (column 5 of Table II) Poisson gave better results, but not sufficiently better for an empirical success to be able to justify an absence of biologically intelligible foundation.

Now let us go to the other extreme and imagine that *all* the cases after the first are due exclusively to personal infection. It has long been orthodox faith that the period of infectiousness is very short, let us conceptually reduce it to an instant and suppose that during an instant of time the m persons are exposed to a constant risk (furnished by the primary case), that if during this instant r are infected, then again for an instant $m - r$ are exposed to risk. We have the conception not of a single binomial distribution but of a *chain* of binomials. Thus in the chosen illustration of $m = 3$ we might reach a final score of three cases not uniquely as in the original use of the binomial through the term p^3, but by any one of the following:

(1) Three cases from the first exposure and of course no more; given as before by p^3.

(2) Two cases from the first exposure and then one case. The probability of the former event is $3p^2q$, of the latter p, because since there is but one person at risk the binomial $(p + q)$ has unity for its exponent. Combined probability, $3p^3q$.

(3) One case from the first exposure and two from the second. The respective probabilities are $3pq^2$ and p^2, the combined probability is $3p^3q^2$.

(4) One case from the first exposure, one from the second and one from the third; the several probabilities are $3pq^2$, $2pq$ and p, their product $6p^3q^3$.

Each term of the original binomial with exponent m is the first link of a

chain of binomials with diminishing exponents, except of course the first and last terms since there can be no more than m cases on the one hand, while on the other, if the initial case fails to infect any of the exposed to risk, the chain breaks at its first link.

Evidently r cases can be generated ($r \not> m$) in as many ways as there are compositions of r (zero not admissible as a part), viz. in 2^{r-1} ways. If $(r_1, r_2 \dots r_s)$ be a composition, we have for the frequency of r_1 primary, r_2 secondary, $\dots r_s$ sth order cases,

$$\binom{m}{r_1} p^{r_1} q^{m-r_1} \binom{m-r_1}{r_2} p^{r_2} q^{m-r_1-r_2} \dots \binom{m-r_1-r_2 \dots r_{s-1}}{r_s} p^{r_s} q^{m-r_1-r_2 \dots -r_s} q^{m-r_1-r_2 \dots r_s}$$

$$= \frac{m!}{(m-r)!\, r_1!\, r_2! \dots r_s!} p^r q^{(s+1)m-[(s+1)r_1+sr_2+\dots 2r_s]} \qquad \dots\dots(1).$$

When m is a small number (1) can be used directly, but my friend Dr Isserlis has devised a much more elegant way of approach.

Let us write

$$\phi_m = p^m + \binom{m}{1} p^{m-1} q \phi_1 + \binom{m}{2} p^{m-2} q^2 \phi_2 + \dots \binom{m}{s} p^{m-s} q^s \phi_s + \dots q^m \qquad \dots\dots(2).$$

This is the required distribution. For $m = 1$, a single exposure exhausts the possibilities, $\phi_1 = p + q$, and there is one link.

When $m = 2$, there may be a second link on the chain:

$$\phi_2 = p^2 + 2pq\phi_1 + q^2.$$

When $m = 3$, we have three possible links:

$$\phi_3 = p^3 + 3p^2 q \phi_1 + 3pq^2 \phi_2 + q^3,$$

and so on.

Now let us expand ϕ_m in powers of p by writing it $\overset{r=m}{\underset{r=0}{S}} A_{m,r} p^r$, where $A_{m,r}$ is a function of q only.

$$(2) \equiv A_{m,m} p^m + A_{m,m-1} p^{m-1} + \dots A_{m,r} p^r + \dots A_{m,0} \qquad \dots\dots(3).$$

In (3) $A_{m,r}$ is the coefficient of p^r; in (2) the coefficient of p^r is: $\binom{m}{m-r} q^{m-r}$ multiplied by the first term of the expansion of ϕ_{m-r} omitting the factor in p, $+ \binom{m}{m-r+1} q^{m-r+1}$ multiplied by the second term of the expansion of ϕ_{m-r+1} omitting the factor in p, $+ \dots \binom{m}{m-r+s} q^{m-r+s}$ multiplied by the $(s+1)$th term of the expansion of ϕ_{m-r+s} omitting the factor in p; hence in the notation of (3)

$$A_{m,r} = \binom{m}{r} q^{m-r} A_{m-r,0} + \binom{m}{r-1} q^{m-r+1} A_{m-r+1,1} + \dots \binom{m}{1} q^{m-1} A_{m-1,r-1} \qquad \dots\dots(4).$$

$A_{m,0}$ is q^m, so (4) enables us to calculate the A's step by step. Thus:

$$A_{m,1} = \binom{m}{1} q^{m-1} A_{m-1,0} = m q^{m-1} q^{m-1} = m q^{2(m-1)};$$

$$A_{m,2} = \binom{m}{2} q^{m-2} A_{m-2,0} + \binom{m}{1} q^{m-1} A_{m-1,0}$$

$$= \binom{m}{2} q^{m-2} q^{m-2} + m q^{m-1} (m-1) q^{2(m-2)}$$

$$= \binom{m}{2} q^{2(m-2)} (1 + 2q^{m-1}).$$

The values up to $A_{m,6}$ are:

$$A_{m,0} = q^m;$$

$$A_{m,1} = \binom{m}{1} q^{2(m-1)};$$

$$A_{m,2} = \binom{m}{2} q^{2(m-2)} (1 + 2q^{m-1});$$

$$A_{m,3} = \binom{m}{3} q^{2(m-3)} (1 + 3q^{m-2} + 3q^{m-1} + 6q^{2m-3});$$

$$A_{m,4} = \binom{m}{4} q^{2(m-4)} (1 + 4q^{m-3} + 6q^{m-2} + 4q^{m-1} + 12q^{2m-5} + 12q^{2m-4} + 12q^{2m-3} + 24q^{3m-6});$$

$$A_{m,5} = \binom{m}{5} q^{2(m-5)} (1 + 5q^{m-4} + 10q^{m-3} + 10q^{m-2} + 5q^{m-1} + 20q^{2m-7} + 30q^{2m-6} + 50q^{2m-5} + 30q^{2m-4} + 20q^{2m-3} + 60q^{3m-9} + 60q^{3m-8} + 60q^{3m-7} + 60q^{3m-6} + 120q^{4m-10});$$

$$A_{m,6} = \binom{m}{6} q^{2(m-6)} (1 + 6q^{m-5} + 15q^{m-4} + 20q^{m-3} + 15q^{m-2} + 6q^{m-1} + 30q^{2m-9} + 60q^{2m-8} + 120q^{2m-7} + 120q^{2m-6} + 120q^{2m-5} + 60q^{2m-4} + 30q^{2m-3} + 120^{3m-12} + 180q^{3m-11} + 300q^{3m-10} + 360q^{3m-9} + 300q^{3m-8} + 180q^{3m-7} + 120q^{3m-6} + 360q^{4m-14} + 360q^{4m-13} + 360q^{4m-12} + 360q^{4m-11} + 360q^{4m-10} + 720q^{5m-15}).$$

The fractional frequency, or probability, that, after a complete exposure, out of m, r will be infected, ${}_mf_r$ say, is merely $A_{m,r}$ multiplied by p^r.

Hence the respective f's can be computed. If $m = r$, (4) becomes

$$A_{m,m} = 1 + \binom{m}{1} q A_{1,1} + \binom{m}{2} q^2 A_{2,2} + \ldots \binom{m}{m-1} q^{m-1} A_{m-1,m-1},$$

and we can compute $A_{1,1}$, $A_{2,2}$, etc., checking the previous calculation.

In actual practice, unless a table of f's for different values of m and p has

been computed, the mean number of cases will be needed; it can be calculated without using A_{mm}. We see from (3) that the mean, which we may denote by $\bar{\phi}_m$, is

$$A_{m,1}p + 2A_{m,2}p^2 + \ldots,$$

i.e.

$$p\phi'_m \text{ if } \phi'_m = \frac{d\phi_m}{dp};$$

from (2) we have

$$\phi'_m = mp^{m-1} + (m-1)\binom{m}{1}p^{m-2}q\phi_1 + \ldots\ldots$$

$$+ \binom{m}{1}p^{m-1}q\phi_1' + \binom{m}{2}p^{m-2}q^2\phi_2' + \ldots\ldots$$

Since $\phi_s = 1$ for all values of s, the first line $= m$. Hence

$$\bar{\phi}_m = \binom{m}{1}pq^{m-1}\bar{\phi}_{m-1} + \binom{m}{2}p^2q^{m-2}\bar{\phi}_{m-2} + \ldots\binom{m}{1}p^{m-1}q\bar{\phi} + mp \qquad \ldots\ldots(2\,a).$$

Now if we write

$$\bar{\phi}_m = B_{m,m}p^m + B_{m,m-1}p^{m-1} + \ldots B_{m,1}p \qquad \ldots\ldots(3\,a),$$

and equate coefficients as before between (2 a) and (3 a), we find that

$$B_{m,r} = \binom{m}{r}q^{m-r}B_{m-r,0} + \binom{m}{r-1}q^{m-r+1}B_{m-r+1,1} + \ldots\binom{m}{1}q^{m-1}B_{m-1,r-1} \qquad \ldots\ldots(5)$$

of the same form as (4). In (4) $A_{m,0} = q^m$, in (5) $B_{m,1} = m$, giving

$$B_{m,2} = \binom{m}{1}q^{m-1}B_{m-1,1} = \binom{m}{1}q^{m-1}(m-1), \text{ etc.}$$

So

$$\frac{B_{m,1}}{A_{m,0}} = \frac{m}{q^m}, \quad \frac{B_{m,2}}{A_{m,1}} = \frac{m-1}{q^{m-1}} \ldots;$$

$$\frac{B_{m,r}}{A_{m,r-1}} = \frac{m-r+1}{q^{m-r+1}}.$$

Hence

$$\bar{\phi}_m = pB_{m,1} + p^2B_{m,2} + \ldots$$

$$= p\frac{m}{q^m}A_{m,0} + p^2\frac{m-1}{q^{m-1}}A_{m,1}, \text{ etc.}$$

These formulae enable us to solve the problem with semi-mechanical effort in any case. Formulae expressed in terms of p and q are not, however, suitable for use. In order to obtain a suitable value of q from the statistics, we require to express the mean of the statistics as a function of the p and q and, that its accuracy may be readily checked, it is better to reduce to terms of q (or p) only. Evidently if the reduction is correct, the sum of the coefficients of powers of q and the constant term must vanish, for if $q = 1$, the mean must be zero. Similarly, the term free of q must be equal to m, for if $p = 1$, the mean must be m. The values of the frequencies and the means for values of m from 2 to 5 are given below.

$m = 2.$

$$f_0 = q^2.$$
$$f_1 = 2q^2(1-q).$$
$$f_2 = 1 - 3q^2 + 2q^3.$$
$$\text{Mean} = 2 - 4q^2 + 2q^3.$$

$m = 3.$

$$f_0 = q^3.$$
$$f_1 = 3q^4(1-q).$$
$$f_2 = 3q^2(1 - 2q + 3q^2 - 4q^3 + 2q^4).$$
$$f_3 = 1 - 3q^2 + 5q^3 - 12q^4 + 15q^5 - 6q^6.$$
$$\text{Mean} = 3 - 3q^2 + 3q^3 - 15q^4 + 18q^5 - 6q^6.$$

$m = 4.$

$$f_0 = q^4.$$
$$f_1 = 4q^6(1-q).$$
$$f_2 = 6q^4(1 - 2q + q^2 + 2q^3 - 4q^4 + 2q^5).$$
$$f_3 = 4q^2(1 - 3q + 6q^2 - 7q^3 + 12q^5 - 21q^6 + 18q^7 - 6q^8).$$
$$f_4 = 1 - 4q^2 + 12q^3 - 31q^4 + 40q^5 - 10q^6 - 56q^7 + 108q^8 - 84q^9 + 24q^{10}.$$
$$\text{Mean} = 4 - 4q^2 + 12q^3 - 40q^4 + 52q^5 - 24q^6 - 60q^7 + 132q^8 - 96q^9 + 24q^{10}.$$

$m = 5.$

$$f_0 = q^5.$$
$$f_1 = 5q^8(1-q).$$
$$f_2 = 10q^6(1 - 2q + q^2 + 2q^4 - 4q^5 + 2q^6).$$
$$f_3 = 10q^4(1 - 3q + 3q^2 + 2q^3 - 6q^4 + 6q^6 + 3q^7 - 18q^8 + 18q^9 - 6q^{10}).$$
$$f_4 = 5q^2(1 - 4q + 10q^2 - 14q^3 + 5q^4 + 16q^5 - 32q^6 + 26q^7 - 20q^8 + 60q^9 - 132q^{10} + 156q^{11} - 96q^{12} + 24q^{13}).$$
$$f_5 = 1 - 5q^2 + 20q^3 - 60q^4 + 99q^5 - 65q^6 - 80q^7 + 205q^8 - 125q^9 + 20q^{10} - 290q^{11} + 820q^{12} - 960q^{13} + 540q^{14} - 120q^{15}.$$
$$\text{Mean} = 5 - 5q^2 + 20q^3 - 70q^4 + 125q^5 - 115q^6 - 60q^7 + 230q^8 - 110q^9 - 80q^{10} - 240q^{11} + 960q^{12} - 1140q^{13} + 600q^{14} - 120q^{15}.$$

Table III. *Frequencies of secondary cases of measles, compared with Chain Binomials (bracketed figures).*

	$m=2$		$m=3$	
0	197 (198·3)	120 (122·4)	84 (89·7)	37 (40·8)
1	104 (101·4)	93 (88·2)	60 (52·7)	34 (27·9)
2	57 (58·3) $P=0·75$	86 (88·4) $P=0·58$	57 (54·4)	42 (42·7)
3	—	—	27 (31·2) $P=0·36$	36 (37·6) $P=0·42$

	$m=4$		$m=5$
0	60 (59·54)	—	25 (26·86)
1	29 (28·87)	—	15 (12·42)
2	25 (24·45)	—	9 (10·63)
3	11 (14·35)	—	10 (6·90)
4	7 (4·79) $P=0·61$	—	1 (3·30)
5	—	—	1 (0·89) $P=0·42$

The sixth column of Table II contains the values deduced on this hypothesis; they plainly agree much better with the observations than the binomial terms.

Table IV.

No. of children under 10 years of age, not having had measles, exposed to infection from a case in same house during 1926 epidemic (St Pancras)

No. attacked (at least 4 days after first case and within a month of last case in house)	1	2	3	4	5	6	7	Total
0	322	120	37	12	3	1	—	495
1	238	93	34	10	4	—	—	379
2	—	86	42	5	5	2	1	141
3	—	—	36	9	9	2	—	56
4	—	—	—	7	2	1	—	10
5	—	—	—	—	1	1	2	4
6	—	—	—	—	—	1	1	2
7	—	—	—	—	—	—	2	2
Total	560	299	149	43	24	8	6	1089
Total children in above houses	560	598	447	172	120	48	42	1987
No. of these attacked	238	265	226	75	54	25	32	915

In Table III the comparison is extended to all values of m up to $m = 5$ and to two other sets of data, for the cases of $m = 2$ and $m = 3$ derived from Table IV (also supplied by Dr Stocks) which relates to exposed to risk somewhat more stringently selected.

Not one of the six values of P (deduced by entering Elderton's table for n' = the number of frequencies less one, since *two* degrees of freedom have been absorbed) is less than 0·35, so that the agreement between fact and fancy is statistically respectable. One might anticipate some improvement by dropping the condition that p is constant throughout the chain, *i.e.* the condition that the danger of a secondary case to the exposed to risk is as great as that of a primary. This condition seems biologically improbable, for we should, especially in view of Dr Stocks' recent work, expect the factor of latent immunisation to play some part. But unless m were larger than it is likely to be in practice, it would be futile to seek to deduce a series of p's one for each remove from the first link. The deduction of each new p would mean the absorption of one more degree of freedom, and we should tend to reach a meaningless, because compulsory, concordance between theory and observation.

I think, speaking as a biological statistician, that this arithmetical device is as successful in describing the facts of measles transmission as we can reasonably expect, and that its application to similar data of scarlet fever, smallpox, diphtheria and whooping-cough would be illuminating. In scarlet fever, one would expect a similar result but with p smaller. Perhaps the ratios of the deduced p's might be an acceptable measure of the infectiousness of scarlet fever in terms of the standard of measles.

In diphtheria the position is more complex, since the period of infectious-

ness of each case (assuming, of course, that the patients are not removed from contact with the exposed to risk) is *not* limited to a short period of time. Under those circumstances the conception of a series of links in a chain is inappropriate. We should, I think, need to return to the conception of a continuous bombardment. As, however, I have no data, speculation is idle. I confine myself to the present, very modest, contribution to the study of infectiousness in diseases appertaining roughly to the measles class.

The arithmetical computation from the formulae given is not difficult. The longest stage is the approximation to q, by the use, for instance, of Newton's approximation. Actually putting $q^m = f_0/F$, where f_0 is the frequency of groups with no secondaries and F the total number of groups, gives a first approximation close enough for one or two applications of the Newtonian approximation to suffice.

Taking as an illustration the most difficult case, we have:

Cases	Frequency
0	25 (26·86)
1	15 (12·42)
2	9 (10·62)
3	10 (6·90)
4	1 (3·30)
5	1 (0·89)
	61

The mean is 1·1803279, so that the equation for q is

$$120q^{15} - 600q^{14} + 1140q^{13} - 960q^{12} + 240q^{11} + 80q^{10} + 110q^9 - 230q^8 + 60q^7 + 115q^6 - 125q^5 + 70q^4 - 20q^3 + 5q^2 - 3{\cdot}8196721 = 0.$$

$25/61 = 0{\cdot}409836$ must be approximately equal to q^5 and 0·836608 is approximately equal to $\sqrt[5]{0{\cdot}409836}$. Hence 0·84 may be taken as a first approximation to the value of q.

Substituting in the equation and in its first differential we have

$$f(x) = -\ 0{\cdot}0877237,$$
$$f'(x) = \ \ 10{\cdot}102029,$$

giving
$$f(x)/f'(x) = -\ 0{\cdot}0086838.$$

A second approximation is 0·84868, and proceeding as before we reach 0·84873 as a sufficiently near approximation for our purpose. Substituting this value of q in the formulae, one reaches the values given in parentheses.

For purposes of *rough* estimation a table of values such as Table V will be found useful. The intervals, however, are too large for it to be possible by interpolation to reach accurate results. Thus, taking the example just worked out, interpolating for the values of the mean by taking 0·5086 of the terms of the series for $q = 0{\cdot}9$ and 0·4914 times the values corresponding in the series for $q = 0{\cdot}8$, one reaches 28·1, 11·7, 9·5, 6·6, 3·8 and 1·3 which are not close approximations to the correct values although, perhaps, sufficient for rough estimations. Since, as already stated, I have no other data than those

provided by Dr Stocks, which reach a respectable standard of accuracy, I cannot myself test the value of the method on the findings in other diseases. But a single, only half-serious experiment, is worth mentioning. Table VI is taken

Table V.

$q=$ ...	·1	·2	·3	·4	·5	·6	·7	·8	·9
$m=2$ f_0	·01	·04	·09	·16	·25	·36	·49	·64	·81
f_1	·018	·064	·126	·192	·25	·288	·294	·256	·162
f_2	·972	·896	·784	·648	·50	·352	·216	·104	·028
Mean	1·962	1·856	1·694	1·488	1·25	·992	·726	·464	·218
$m=3$ f_0	·001	·008	·027	·064	·125	·216	·343	·512	·729
f_1	·0003	·0038	·0170	·0461	·0938	·1555	·2161	·2458	·1968
f_2	·0248	·0829	·1561	·2281	·2813	·2972	·2620	·1751	·0637
f_3	·9739	·9052	·7999	·6618	·5000	·3313	·1790	·0671	·0105
Mean	2·9717	2·8854	2·7289	2·4877	2·1563	1·7437	1·2769	·7974	·3557
$m=4$ f_0	·0001	·0016	·0081	·0256	·0625	·1296	·2401	·4096	·6561
f_1	·0000	·0002	·0020	·0098	·0313	·0746	·1412	·2097	·2126
f_2	·0005	·0062	·0251	·0624	·1172	·1782	·2186	·1990	·0968
f_3	·0301	·0939	·1686	·2396	·2891	·2944	·2385	·1315	·0297
f_4	·9693	·8981	·7961	·6626	·5000	·3232	·1616	·0502	·0049
Mean	3·9685	3·8866	3·7427	3·5037	3·1328	2·6069	1·9403	1·2030	·5147
$m=5$ f_0	·0000	·0003	·0024	·0102	·0313	·0778	·1681	·3277	·5905
f_1	·0000	·0000	·0002	·0020	·0098	·0336	·0865	·1678	·2152
f_2	·0000	·0004	·0036	·0155	·0439	·0940	·1567	·1908	·1229
f_3	·0007	·0084	·0307	·0707	·1257	·1829	·2103	·1646	·0527
f_4	·0343	·0999	·1724	·2410	·2893	·2915	·2255	·1080	·0161
f_5	·9649	·8910	·7906	·6606	·5000	·3203	·1530	·0412	·0026
Mean	4·9641	4·8804	4·7422	4·5120	4·1321	3·5376	2·6975	1·6811	·6966

Table VI.

Households having cases ...	1	2	3	4	5	6	7	Total
No. of households	234	101	57	30	12	4	2	440
No. of persons	1083	513	326	186	78	32	19	2237
Average no. of persons per house	4·63	5·01	5·72	6·2	6·5	8	9·5	—

from the *Ministry of Health's Report on the Pandemic of Influenza*, 1918–19. I arbitrarily simplified the table and assumed that it referred to groups of five in family and would thus read:

Secondary cases	No. of families
0	234
1	101
2	57
3	30
4	12
5	6
	440

The mean is 0·87. Treated as a binomial with exponent 5 we should have $p = 0·174$ and $q = 0·826$. This would yield as frequencies 169·2, 178·2, 75·1, 15·8, 1·7 and 0·1. Clearly preposterous values. The Poisson frequencies are 184·5, 160·2, 69·7, 20·3, 4·4 and 0·9 (grouping 5 and onwards), almost equally preposterous. Interpolating for a "chain" we have 239·4, 91·0, 59·3, 31·9, 14·2, 4·1; a quite reasonable fit and, as we have seen, this is not the best fit which we could obtain (the experiment is too trivial to justify the labour of

careful fitting). One may say that the influenza of 1918–19 seems to have behaved rather like measles in its domestic evolution. However, the object of this paper is to bring to the notice of others a method, not to discuss, without data, epidemiological results.

To Dr Isserlis I owe not only the elegant demonstration quoted but helpful criticism of the whole idea and Mr W. J. Martin has been indefatigable in sparing me laborious arithmetic. To our valued colleague the late Mr H. E. Soper, I owe the painless destruction of various other weakly dream children brought to birth during this study and the encouragement that this brat might be worth preservation.

REFERENCES.

GREENWOOD, M. and YULE, G. U. (1920). *J. Roy. Stat. Soc.* **83**, 255–79.
MAYNARD, G. D. and TROUP, J. McD. (1911). *Biometrika*, **8**, 396–404.
NEWBOLD, E. M. (1927). *J. Roy. Stat. Soc.* **90**, 487–535.
PEARSON, K. (1911). *Biometrika*, **8**, 405–12.
—— (1913). *Ibid.* **9**, 28–33.

(*MS. received for publication* 2. XII. 1930.—Ed.)