

MISCELLANEA.

I. On the Distribution of Severity of Attack in Cases of Smallpox.

By KARL PEARSON, F.R.S.

Dr Turner finds that the protection provided by vaccination does not diminish as rapidly with age as he would expect it to do. His expectation arises from the fact that current medical opinion considers that the immunity provided by vaccination diminishes rather rapidly with the increase of the period which has elapsed since vaccination. I do not propose to consider whether protection against death when disease is incurred is really of the same character as immunity against an attack. But I should wish to point out that Dr Brownlee has reached a very similar conclusion to Dr Macdonell on this point by a very different process, and further that it follows, if we simply use the coefficient of association, which makes no appeal to normality of distribution. On the other hand, while I am distinctly interested in Dr Turner's theory of a "curtailed normal curve," I am compelled to say that I consider it an extremely improbable hypothesis in the present application. The group of persons who catch smallpox are a *selected* group of the general population, selected because (i) they have come in contact with the disease, (ii) they were at the time of such contact not sufficiently immune against it. This sufficiency of immunity must depend not only on prior vaccination, but on a host of other causes, the virulence of the poison they encountered, their particular state of health at the time in question, their conduct before and after the risk was run, etc., etc. It is impossible to suppose a rigid line drawn at a certain grade and say all below this grade escape this disease, all above will contract it. I cannot understand how those 'selected' to incur the disease differ from any other naturally selected group with which we are acquainted. Now biometricians are dealing with selected groups every time they measure the variation of a character in a species, but no such truncated normal distributions have yet exhibited themselves*.

Dr Turner says "A curtailed table in measurable characters would result if we collected statistics of height in soldiers, a population from whom all below a certain standard height had already been rejected" (p. 497 above). Now this is a case which can be well tested. For example, height standards exist for both the American and Italian armies. Yet what do we find? That the distribution of the statures of the accepted recruits in both countries, so far from forming a curtailed normal distribution is in some cases as close to a complete Gaussian

* American trotting horses give the nearest approach, there being a time limit to entry in the record of trotters. But this is a perfectly arbitrary line drawn across the trotting population by the hand of man, and not a selection due to a complex of natural causes.

curve as anything with which we are acquainted*. The fact is that there are so many other factors on which selection depends that we get no "curtailing" of the distribution at all. I should imagine that it would be precisely the same with those who are selected to incur small-pox. The selection depends on so very much else than a certain absolutely fixed grade of immunity. This point is of such importance that I think it worth illustrating by an example. It is one which Professor Weldon has most kindly provided me with out of the extensive reductions he has made of Italian conscripts and recruits. In the Province of Verona in the five years between 1875 and 1879 there were 16,203 conscripts† and from these were selected 3810 recruits‡. The following table contains the two distributions:

Stature in Centimetres	184-125	186-127	128-129	130-131	132-133	134-135	136-137	138-139	140-141	142-143	144-145	146-147	148-149	150-151	152-153	154-155	156-157	158-159	160-161
Conscripts	4	—	2	1	1	1	—	2	7	7	22	33	55	117	225	626	725	1027	1503
Selected Recruits	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	16	148	239	354

Stature in Centimetres	162-163	164-165	166-167	168-169	170-171	172-173	174-175	176-177	178-179	180-181	182-183	184-185	186-187	188-189	190-191	192-193	Totals
Conscripts	1869	2065	2125	1703	1525	1058	699	384	205	131	45	24	7	4	1	—	16203
Selected Recruits	471	542	538	425	414	256	170	133	55	35	9	3	—	2	—	—	3810

In the diagram the two distributions with their corresponding curves are given, and we see that in the selected recruits there is not the slightest approach to a curtailed normal curve. In fact, if we examine the fundamental constants of the distribution§ we find:

For the Conscripts

$$\sqrt{\beta_1} = .1658 \pm .0130$$

$$\beta_2 = 3.7516 \pm .0260$$

For the Recruits

$$\sqrt{\beta_1} = .3409 \pm .0268$$

$$\beta_2 = 2.7989 \pm .0536.$$

The probable errors are those which would arise if the distribution were truly normal¶. The deviations from symmetry are for the two cases 12.6 and 12.7 times their probable errors respectively. It is impossible therefore to say more than that the two curves are both markedly skew and deviate equally from normality in this respect. In the next place the deviations from mesokurtosis are in the two cases 28.9 and 3.8 times their probable errors; in other words the curve for the conscripts diverges indefinitely more from the normal curve than that for the selected recruits. This is only one instance out of many which emphasise the same points, i.e. that there is no approach in the selected curve to curtailment and it differs on the whole less from normality than is the case with the unselected material.

* See, for example, Baxter's statistics for U.S. recruits discussed, *Phil. Trans.* Vol. 186, A, p. 385.

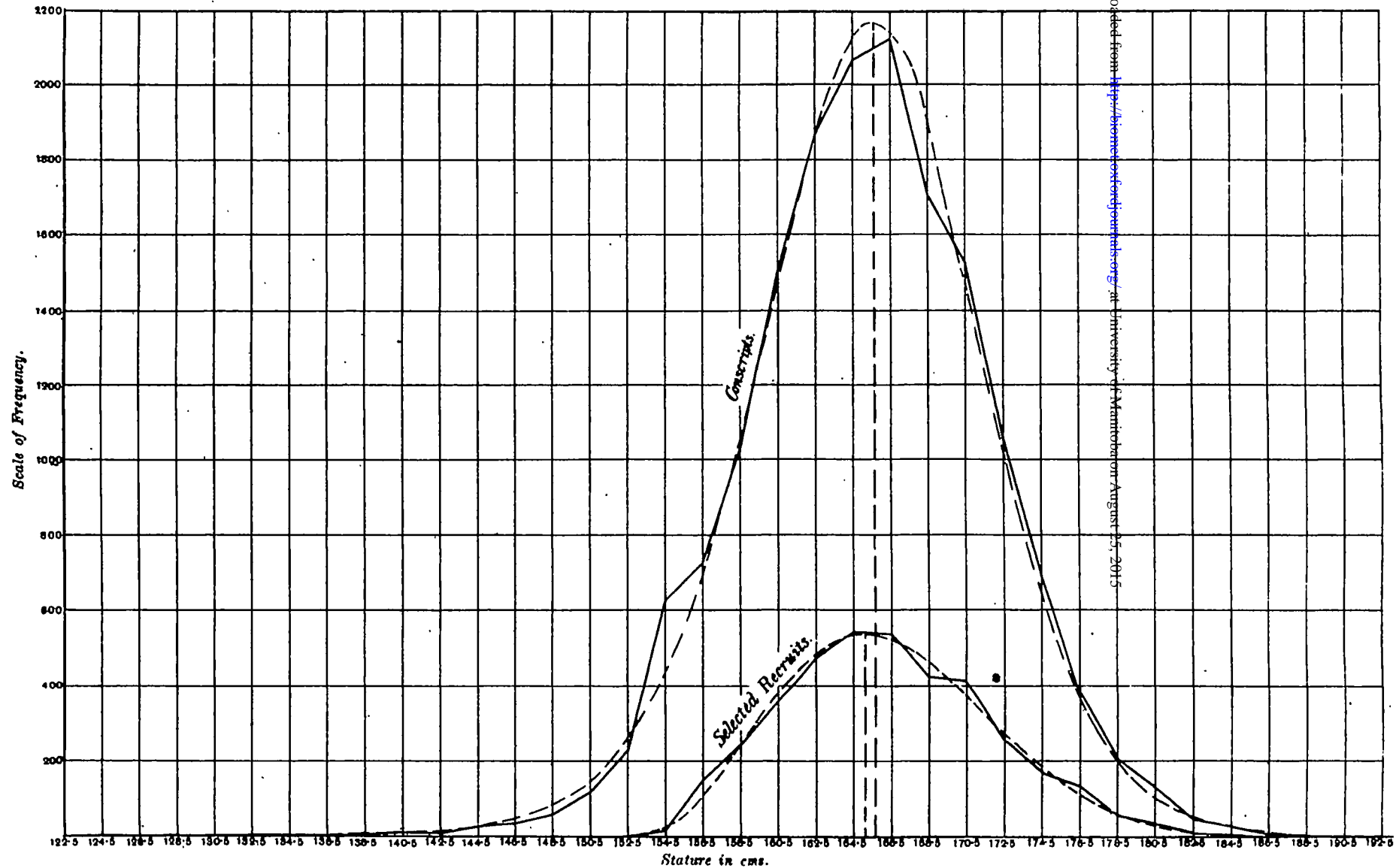
† R. Livi, "Classificazione delle stature.....dei coscritti delle leve di terra negli anni 1875-79." *Annali di Statistica*, Vol. VIII. 1883, pp. 144-9.

‡ R. Livi, *Antropometria Militare*, Parte I., p. 280, Roma, 1898.

§ *Phil. Trans.* Vol. 186, A, p. 368. *Biometrika*, Vol. IV., pp. 174-7.

¶ *Phil. Trans.* Vol. 198, A, p. 278.

Comparison of Frequency Distributions of Verona Conscripts and Selected Recruits.



I have next endeavoured to form an estimate of the distribution of severity in the cases of smallpox, and the only method of doing this seemed to be to form a frequency distribution for the intervals which elapse between either (a) onset or (b) eruption and the first bath. This may be taken in a rough sort of way as a measure of the severity of the attack. Unluckily any character which depends on time usually gives a skew frequency distribution and the present case is no exception to the rule. But I do not think that the resulting distributions can be in any way described as curtailed normal curves. Dr John Brownlee kindly provided me with the particulars of between 800 and 900 cases. I could distinguish no sensible differences in the male and female distributions. Further, the vaccinated cases formed a very large proportion of the total, for example 779, as against 55 unvaccinated and 21 doubtful. Hence, taking into account that what Dr Turner is dealing with is the distribution of cases in all classes, vaccinated or unvaccinated, I have clubbed all groups together to get my distribution of severity. There were 57 deaths, which must of course be excluded, from a time to bath test of severity, they occurred with rather an erratic distribution at a mean interval of 10·4 days after onset or 7·6 days after beginning of the eruption. Thus the interval between onset and eruption in the fatal cases is about 2·8 days, while in the recoveries it is 3·1 days, so that there may, when more material is forthcoming, be found to be a sensible difference in this interval for the two classes of cases.

The following table gives the distribution of severity of attack as measured by the two tests of days (i) from onset to bath and (ii) from eruption to bath.

Days	4-5	6-7	8-9	10-11	12-13	14-15	16-17	18-19	20-21	22-23	24-25	26-27	28-29	30-31	32-33	34-35	36-37	38-39	40-41	Totals
Onset to Bath ...	2	13	40	131	192	152	99	73	40	24	13	17	10	6	4	6	2	1	—	826
Eruption to Bath	10	55	164	174	156	96	71	39	28	13	21	6	8	2	7	3	3	1	—	855

Neither of these distributions can be considered as a curtailed normal curve. They must, I think, be looked upon as significantly skew distributions of the usual type such as in practice almost invariably arise, when time is the variate to which we plot frequency. They do not appear to me to give any support to Dr Turner's view expressed on p. 496 above as (2), or to the suggestion on p. 497, that the mildest cases are more frequent than those more severe; the different degrees of severity do not diminish without exception as the degree increases, and since the bulk of our cases are all vaccinated, it is not possible to suppose that the actual frequency curve among the vaccinated is of the type suggested by Dr Turner. The modal frequency corresponds to a sensible degree of severity of attack, i.e. to about 13 days from onset to first bath, while the mildest cases correspond to only 4 days. It seems therefore quite impossible to suppose, as Dr Turner does on p. 502, that the severity frequency distribution is half a normal curve of all exposed to risk. The distribution of the attacked very considerably passes the mode, and on the assumption made by Dr Turner, those exposed to risk and escaping must be a very small fraction indeed of the total population exposed to risk. Indeed the above distributions show that the severity of the attack rises from zero to a maximum, and then falls, in the usual skew frequency fashion, at a slower rate to zero. The constants of the above distributions have been worked out; it will suffice here to give, however:

$$\text{Onset to Bath} \quad \beta_1 = 2.3229, \quad \beta_2 = 6.2466.$$

$$\text{Eruption to Bath} \quad \beta_1 = 2.606, \quad \beta_2 = 6.3664.$$

Thus both distributions are sensibly non-Gaussian.

It may be said that this skewness might itself account for the results which Dr Turner considers opposed to current medical opinion and which he would account for by a curtailed normal distribution. Possibly it may, but I hold that this is very unlikely, and for the following reasons:

(i) The severity is here measured by a *time* test, which invariably gives a skew distribution; but if severity could be measured in some other quantitative manner, for example by intensity of eruption, mean rise of temperature or in some such fashion, we should most probably reach a more normal distribution.

(ii) The effect of even the large amount of skewness indicated above on the correlation found by a fourfold division is, I think, not likely to be very large. It is too often assumed that the distribution must be *very closely* normal to give a good result by the fourfold table method. As a matter of fact, if the fourfold division falls on the "long tailed side" of the mode, as it does in these vaccination statistics, we get close values to the actual correlation even by the fourfold table method. To illustrate this I took the only correlation which was available on the data, namely the correlation between onset to eruption and eruption to bath. This correlation was found by the ordinary product moment process, which is independent of any hypothesis of normal distribution and came out $= -\cdot174 \pm \cdot023$. In other words, if severity of the disease be measured by a long period from eruption to bath, then a short period between onset and eruption is associated with severity. This is in accordance with the previous indication that the fatal cases have a short period from onset to eruption.

I now formed a fourfold table from my exceedingly skew distribution, dividing between 14 and 15 days from eruption to bath and between 3 and 4 days from onset to eruption. This gave:

Onset to Eruption.

Eruption to Bath.		3 days and less	4 days and more	Totals
	14 days and less ...	395	235	630
	15 days and more ...	138	51	189
	Totals ...	533	286	819

whence the correlation $-\cdot168$ was deduced. It is clear that the skewness has not sensibly influenced the value of the correlation as determined from this fourfold table.

The divergence of the fourfold dividing lines from the median is certainly not quite as great in this case as in the vaccination problem, but it is difficult to go nearer the tails and get anything like a reliable result. Dividing between 4 and 5 days from onset to eruption, and 16 and 17 days from eruption to bath, I get:

Onset to Eruption.

Eruption to Bath.		4 days and less	5 days and more	Totals
	16 days and less ...	595	104	699
	17 days and more ...	112	8	120
	Totals ...	707	112	819

In this case the probable error of the frequency 8 is no less than 2, but a change of 2 in the value of this frequency would change the value of r by about 20 p.c. Such a division is

therefore very unreliable*. We find, however, that the correlation = $-.238$ with a probable error of about $\pm .05$. In other words within the limits defined by the probable errors, we get a correlation sensibly equal to the actual value $-.174$. I think accordingly that we may assume that the skewness of our severity test would not materially alter the condition of affairs.

I conclude as follows :

(i) There is no reason *a priori* for supposing that the distribution of severity of attack in smallpox follows a curtailed normal curve. Such curves are contrary to any existing experience of the distribution of frequency in *selected* groups. In particular they do not occur in the case of soldiers suggested by Dr Turner.

(ii) The only tests I have been able to apply to the frequency distribution of severity is that of length of period from onset or eruption to bath. These both give a continuous and not a curtailed curve. The curves are skew, as we are accustomed almost invariably to find them when time is the variate.

(iii) This skewness, due to the time as variate, may not really be characteristic of the distribution of smallpox severity; this could only be determined if we chose a variate other than time to measure severity. But if it were characteristic, it does not appear that a fourfold division taken so far towards the tail as occurs when we divide severity at death would seriously affect the result. It is shewn that within the limits of probable error the product moment method of calculating the coefficient of correlation and the fourfold division method lead to sensibly the same results, even when we have a skewness in the distribution as great as is indicated in the eruption to bath test of severity.

II. On the Mean Duration of Life of Individuals Dying within a Year after Birth†.

By RAYMOND PEARL, Ph.D.

The usual custom in tabulating census returns of mortality at different ages is to use a five year base unit for ages above 5, and below that age a one-year unit. This method of tabulation makes the finding of the moments of the frequency distribution somewhat less simple than would be the case if all the base elements were equal. Furthermore, the age distribution of the heavy mortality of the first year of life is not given at all. It becomes a very important matter to know, at least approximately, this first year distribution when one attempts to find the moments for the whole material, because the frequency in this element must be centred at some point before one can proceed with the calculations. The mean age at death of those dying under one year must be known. In connection with some work on vital statistics which is being carried on in the Zoological Laboratory of the University of Michigan, it became necessary to have as exact a determination of the centering point of this first year mortality as it was possible to get. It is the purpose of this note to present the results obtained.

* Dr Turner's Tables XIV. and XV. contain only 8 and 2 cases in their fourth quadrants and the probable errors of the resulting coefficients are very considerable. I think Dr Brownlee's figures (*Biometrika*, Vol. iv. p. 825; *Journal of Hygiene*, Vol. v. pp. 533—4) may be safely taken as giving a reasonable measure of the reduction with time of vaccination protection.

† Contributions from the Zoological Laboratory, University of Michigan, No. 88.