SCIENCE

VOL. LVIII AUGUST 10, 1923 No. 1493

CONTENTS

The Statistical Significance of Experimental Data: PROFESSOR EDWIN B. WILSON 93

The Selection of Subjects for Research: PROFESSOR EUGENE C. BINGHAM 100

Scientific Events:

The British Journal of Experimental Biology; Fellowships in Medicine; The Influence of Modern Science on History and Civilization; Explorations for Rubber; Scientific Work in Siberia; Appointments in Agriculture at the University of California 102

Scientific Notes and News 104

University and Educational Notes 106

Discussion and Correspondence:

Proposals for the Preservation of the Wisent: DR. THEODOR G. AHRENS. *The American Educator and Scientist:* F. C. CALKINS 107

Scientific Books:

Grundzüge der Paläontologie: DR. W. D. MATTHEW 107

Quotations:

The Annual Exposition of Chemical Industries 109

Special Articles:

The Production of "Brown-Séquard's Epilepsy" in Normal Non-operated Guinea Pigs: DR. JOHN AUER 109

Science News viii

SCIENCE: A Weekly Journal devoted to the Advancement of Science, edited by J. McKeen Cattell and published every Friday by

THE SCIENCE PRESS

Lancaster, Pa. Garrison, N. Y.

New York City: Grand Central Terminal.

Annual Subscription, $6.00. Single Copies, 15 Cts.

SCIENCE is the official organ of the American Association for the Advancement of Science. Information regarding membership in the association may be secured from the office of the permanent secretary, in the Smithsonian Institution Building, Washington, D. C.

Application made for transfer of entry as second-class matter to the Post Office at Lancaster, Pa.

THE STATISTICAL SIGNIFICANCE OF EXPERIMENTAL DATA[1]

WHEN a few days ago your secretary, Mr. W. T. Bovie, acting on pressure from your chairman, Mr. J. S. Hughes, urged me with their well-known energies to speak in this symposium they left me little chance to refuse. As I understand the circumstances I am a sort of "pinch hitter" for Mr. J. Arthur Harris, whose long-continued biometric studies would clearly indicate him for this place, but whose absence in the south made it necessary to find a substitute. From him you might reasonably have expected a home run; you must be content with me if I bunt out a one-bagger just to keep the game going.

I should have liked to have more time for preparation. The literature upon the statistical aspects of feeding experiments is not microscopic and the data available for statistical study are extensive. Moreover, my Yale training received here twenty odd years ago under J. Willard Gibbs was not such as to make comfortable for me the presentation of somewhat hastily collected notes. There was not in those days the fervid impatience in science that has developed in recent times in some quarters, and Gibbs himself was a model to any young man not only in his scientific thinking but in his modest and painstaking contemplation of some of the most intricate problems of nature—statistical problems. It may not be amiss if I quote these words from the preface of his last great work entitled "Elementary Principles in Statistical Mechanics" written in 1901:

We avoid the greatest difficulties when giving up the attempt to frame hypotheses concerning the constitution of material bodies, we pursue statistical inquiries as a branch of rational mechanics. In the present state of science, it seems hardly possible to frame a dynamic theory of molecular action which shall embrace the phenomena of thermodynamics, of radiation, and of the electrical manifestations which accompany the union of atoms. Yet any theory is obviously inadequate which does not take account of all these phenomena. Even if we confine our attention to the phenomena distinctly thermodynamic, we do not escape difficulties in as simple a matter as the number of degrees of freedom of a

[1] An address prepared by request as part of a symposium on feeding experiments held by the Biochemical Section of the American Chemical Society, meeting in New Haven during the week of April 2–7 in connection with the dedication of the new Sterling Chemical Laboratory of Yale University.

diatomic gas. . . . Certainly, one is building on an insecure foundation who rests his work on hypotheses concerning the constitution of matter.

Difficulties of this kind have deterred the author from attempting to explain the mysteries of nature, and have forced him to be contented with the more modest aim of deducing some of the more obvious propositions relating to the statistical branch of mechanics. Here, there can be no mistake in regard to the agreement of the hypotheses with the facts of nature, for nothing is assumed in that respect. The only error into which one can fall is the want of agreement between the premises and the conclusions, and this, with care, one may hope, in the main, to avoid.

How very antiquated this sounds to-day when for much of the past two decades the members of the American Chemical Society have been listening as they have this very week to numerous attempts to frame hypotheses concerning the constitution of material bodies, however insecure a foundation such hypotheses may have appeared to Gibbs. The mysteries of nature which Gibbs left somewhat to one side as beyond his modest aim were, however, the mysteries of inorganic nature. Yet in this section of your society we face the vaster difficulties of organic nature, of living matter, and in this symposium the mysteries of *nurture*. It is doubtful if we may hope even with all due care to avoid, in the main, a falling into error. We are confronted with variability of feeds, however carefully we may try to uniformize them, with variation in the experimental animals, however so carefully we select them, with inadequacy of statistical material, no matter how diligently we collect the data. It is necessary to bring to bear every possible check, to exercise all conceivable care in judgment, and yet withal to be modest in our conclusions.

One check, one basis for judgment which we have to-day we did not have in readily available form a quarter century ago. The statistical method in biometrics was just then gaining headway at the hands of Pearson and his incipient school. I remember very clearly with what interest and instruction I read those early papers of Pearson while studying the statistical method with a brilliant young Yale economist in the years 1899–1902. They seemed to open up large possibilities for a scientific basis of political economy quite different from that other scientific method followed by Walras, by Pareto and by Irving Fisher in one of his notable early papers. On the whole, however, it is my judgment that up to the present time it has been not the economist but the student of biology who has most availed himself of these newer statistical methods; it would be interesting to ponder the reason why.

I have been asked to define certain terms and illustrate certain calculations employed in the statistical treatment of experimental data, and this I shall presently do. But in the first place I wish to make some general comments on the philosophy of the statistical interpretation of experimental data. A method is a dangerous thing unless its underlying philosophy is understood, and none more dangerous than the statistical. Our aim should be, with care, to avoid, in the main, erroneous conclusions. In a mathematical or strictly logical discipline the care is one of technique; but in a natural science and in statistics the care must extend not only over the technique but to the matter of judgment, as is necessarily the case in coming to conclusions upon any problem of real life where the complications are great. Over-attention to technique may actually blind one to the dangers that lurk about on every side—like the gambler who ruins himself with his system carefully elaborated to beat the game. In the long run it is only clear thinking, experienced feeling and a patient poise, not automatic systems and methods, that win the strongholds of science—witness the lives and works of those founders of two branches of chemistry of prime importance to this biochemical section: Gibbs in physical chemistry and Pasteur in the chemistry of life.

If you undertake to measure a room for wallpaper, or a court for tennis, you take some simple measuring device and proceed. You determine the measures needed. You may repeat the work as a check but not for the purpose of averaging the results. There is nothing statistical in your mind. The same is true for all our ordinary weights and measures; we weigh and measure; we may check or get somebody else to check the result, we do not average. When a finer or more accurate measure is needed we have recourse to a better or more sensitive instrument; if none can be had with sufficient sensitivity we may even resort to devising one. Generalizations are unsafe, but I will venture the guess that it is always our ideal to have at hand instruments that will enable us directly to read the measures desired and thus spare us the statistical method of analysis. The history of the development of physics has been constantly attended by the conflict between the more accurate instrument and the call for the ever more precise determination of physical properties of matter.

But the time seems always to come when the necessary precision transcends the available appliance. This time came early in the refined measures of astronomy, and already a hundred years ago the treatment of astronomical data was statistical. The most careful measures were most carefully and patiently repeated and the resulting mass of material was reduced by the method of averages and of least squares. The errors or departures from the mean were all small, for the simple reason that the observational

methods and the objects observed were of such a nature that a high degree of precision was directly attainable. Such small deviations as remained to be reduced by statistical treatment were due to a large number of forces or causes each of which if operating alone might produce a considerable irregular and asymmetric variation in the observations, but which were balanced in such a way that the actual deviations were not only small but had a high degree of symmetry and lawfulness. It was under such circumstances that the so-called normal law of errors, proposed and discussed by Laplace about the time that Gauss was learning to walk, came into general use. The law is often called Gauss's law and is figured geometrically as you all know by the bell-shaped probability curve (Fig. 1).

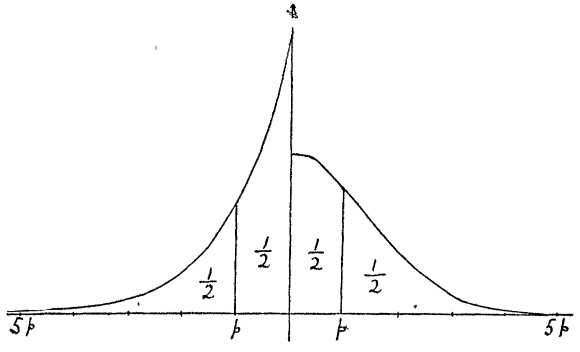

FIG. 1. On the left is Laplace's frequency curve; on the right is Gauss's. The curves are here plotted to a scale which makes the actual probable error p the same; the areas under the two curves are equal and each is divided into two equal parts by the vertical ordinate at p. Each curve should be reproduced symmetrically on the other side of the central vertical line, but the drawing is clearer when only one half of each curve is shown.

During the past century and a fraction a large number of alleged proofs of the normal law of errors or deviations from the arithmetic mean have been proposed and a variety of theoretical objections to it have been propounded. Bertrand brushes aside alike proof and objection with the statement: "Gauss's formula should be adopted. Observation confirms it; that is sufficient for its applications. Its consequences minutely examined are always found accordant with the facts." He is writing, be it clearly understood, from the viewpoint of the astronomer or physicist engaged in precise measurements many times repeated. Indeed, under these conditions we must even recognize the validity of theoretical criteria which may be applied to any such set of data to determine whether or not they are bona fide or have been retouched. Moreover, it is astonishing to find by experience how few observations will serve as a large number. I well remember an experiment I performed with a mirror galvanometer. Some 15 readings were taken to establish a mean. As is always the case, the galvanometer was moving somewhat erratically, due to a large number of conflicting forces—air currents, changes of temperature, electric cars passing in the distance, etc.—despite an effort carefully to shield the vitals of the instrument from external influences. Some of the readings were seemingly unduly far from the mean, and yet if they were discarded the results would not so well accord with the theoretical checks and would tell the tale of presumptive experimental dishonesty. It would have been interesting to apply such criteria to Dr. Cook's famous polar observations; it would take a deal of trouble so to fake the observations as to avoid detection.

So excellent were the results of Gauss's law that many came to believe that it was of wider applicability than either its philosophical premises or its experimental verifications warranted. One of the triumphs of modern statistical theory has been the return to the more general consideration of Laplace and our emancipation from the tyranny of a law too restricted to serve in biometrics. True, there are some fields in which a close observance of the normal law is revealed, as in the distribution of heights among the members of a large population. But little as we know about the true cause of the specific height of any one of us we do know that many major, more or less independent, causes are at work, that these are balanced effectively about a mean from which the deviations are reasonably small. There is no obvious reason why the distribution of Gauss should not apply.

The matter is very different if we study the length of our lives instead of that of our bodies. The frequency distribution of the ages at death among a large population is a reasonably definite affair of undoubted major biological significance; but the curve is not symmetric and has little similarity to that of Gauss. According to the American Experience Mortality Table starting with 100,000 persons alive at the age of 10, the number of deaths per annum is about 750 the first year, decreases slowly to about 720 at the age of 27, increases to about 1,000 at 50 years, and then rapidly to a maximum of 2,500 at the age of 73, from which it falls rapidly to zero in the next 25 years. Now the length of our life surely depends, like the length of our bodies, upon a large variety of conflicting causes, but they do not produce a symmetrical balance about a mean nor are the deviations from the mean small in the sense that they are in precise physical observations or even in the case of our stature. The law of Gauss is plainly contra-indicated.

Although many of the frequency distributions of life are asymmetric and although the general theory

of curve fitting is of importance in biometry and will in due time become important even for such elementary matters as the statistical study of data resulting from feeding experiments, we shall restrict ourselves here to symmetrical laws wherein the chances of positive deviations are equal to the chances of negative deviations of the same magnitude. And of the symmetric laws other than Gauss's one of the most interesting is the first law of errors or small deviations ever proposed. It was put forward by Laplace in 1774 four years before he suggested the normal law. Let us measure deviations not from the arithmetic mean but from the median, that is, from that datum which stands in the middle of a series of observations arranged according to magnitude. The median or middle datum is much used in economic studies instead of the arithmetic mean and in many ways is simpler to use. Now if d be the deviation from the median, Laplace's suggestion is that the frequency of occurrence of a deviation of magnitude d is proportional to

$$e^{-kd} = \frac{1}{(2.718\ldots)^{kd}}$$

In the figure we have on the right the normal law and on the left the law of Laplace. It should be observed that each of the curves should be reproduced symmetrically on the other side of the vertical line, because both are symmetric; but this would complicate the figure. Note that both laws make small deviations much more frequent than large ones, as is ordinarily the case in symmetrical laws. But the Laplace law begins to fall off rapidly and falls off steadily more slowly; whereas Gauss's begins by falling off slowly, then falls faster and finally falls slowly again. The relative numbers of deviations which lie between definite limits are measured by the area under the curve between verticals; and if a vertical be drawn at such a point that the whole area to one side of the middle line is cut exactly in two the deviation represented by that position is called the *probable error* or *probable deviation* p. This is a technical term which in common parlance means merely that the betting is even as to whether a deviation will be greater or will be less than the probable deviation—that in the long run there will be as many deviations larger as there are smaller than this. The curves have been drawn in such a manner that the probable deviations are equal.

As the curves are constantly falling off with increasing deviations, the areas under the curves, toward the right or toward the left as the case may be, are also diminishing and the chances of really large errors are very small. It is customary to take as the unit of reference the probable deviation. The following table shows the chance of deviations being greater than 1, 2, 3, 4 or 5 times the probable value p.

TABLE I

Chances of a deviation greater than 1, 2, 3, 4, 5 times the probable

Laplace's Law		Gauss's Law	Tchebycheff's Criterion
$1p$	.5000	.5000	.....
$2p$	.2500	.1773	less than 0.550
$3p$	.1250	.0430	" " 0.233
$4p$	.0625	.0070	" " 0.137
$5p$	.0312	.0007	" " 0.088

This table shows how fast the chances of large deviations diminish; it shows further how much faster they diminish under the normal law (Gauss) than under Laplace's first law. Only about seven observations in 1,000 can be greater than four times the probable deviation on the so-called probability law; whereas 62 observations in 1,000 may be greater on Laplace's law.

I desire to lay some stress on this table and the inferences from it, because biometricians test the statistical value of a magnitude by reference to the size of its probable error. Statistically determined magnitudes are written followed by a plus or minus sign ($\pm$) and by their probable errors. Thus $x = 12.73 \pm 0.27$ means that the quantity x has been determined by statistical processes, such as averaging, to have the value 12.73 and that the probable error of the determination is 0.27. Or it is even betting that the true value of x lies between $12.73 - 0.27 = 12.46$ and $12.73 + 0.27 = 13.00$. If now we have evidence that the law of the frequency of the distributions is Laplace's second law (the law of Gauss), we may go further and say that the chances are only 43 in 1,000 that the quantity x lies outside the limits defined by thrice the probable error:

$$12.73 - 3 \times (0.27) = 11.92 \text{ and } 12.73 + 3 \times (0.27) = 13.54.$$

Odds of $1{,}000 - 43 = 957$ to 43 or better than 22 to 1 are so high that they represent a reasonable degree of certitude and consequently, when the probable error of a magnitude is less, particularly when it is much less, than one third of the magnitude itself we conclude that the magnitude is statistically significant.

Such a conclusion for its cogency depends tacitly on the fact, if it be a fact, that the chances of large deviations fall off very rapidly with the increase in the deviation as is the case with Gauss's law. On Laplace's law there is not merely one chance in 23, there is one chance in eight that an error exceeds thrice the probable error. If the frequency distribution is unknown, the chances of large deviations are likewise unknown and there is no safe theoretical ground on which those chances can be estimated by

the values set down under Gauss's law in the table. There is a theorem due to the famous Russian Tchebycheff which states that the error in an average value does not exceed a multiple mp of the probable error oftener than once in $2.2/m^2$ times. In the table I have inserted Tchebycheff's Criterion to show how pessimistic he is as to the certitude of statistical inferences in comparison with Gauss. Why is it, then, that biometricians, who deal with material often scanty and of great complexity and diversity of law and with errors neither small nor symmetric, place such confidence in the probable error that a deviation of thrice the probable is regarded as almost impossible and a deviation of four times the probable as quite impossible?

Adequately to answer this question would require an elaborate behavioristic study of biometricians, and I fear that as laboratory animals they would be such varied and variable material that the probable errors of the results would be comparable in magnitude with the results themselves, so that any statistical inference in answer to the question would be illusory. But you can not treat a scientific man or his work exclusively or even largely by the statistical method; the question of judgment must be considered. A scientific investigator, particularly one of the leaders, develops a feeling for his work, an experience in it, and a judgment often so sound that his conclusions merit our most respectful confidence even when those conclusions are apparently founded on very little else than the investigator's intuition. The mature "hunch" of a genius is better than many a scientific demonstration. And it is undoubtedly the experience and belief of many statisticians of the first water that somewhere between three and four times the probable error comes the safe point in drawing conclusions. I mean the safe point for them.

Let me tell you a story of the great astronomer William Herschel. He desired to know the direction of the sun's motion in space. This is a statistical problem; for by the motion of the sun in space is meant its motion relative to the rest of the stars, of which there are many millions visible to our telescopes. Now Herschel had a keen personal acquaintance with the stars and selected with great judgment just seven with respect to which he would determine the solar motion. It would be a rash young statistician who would maintain that seven stars were enough to constitute a fair sample of the sidereal universe for any statistical purpose. Yet Herschel determined the direction of the solar motion with what has proved to be marvelous exactness. We may allow something for luck, but we must not too much discount the efficacy of a Herschel's intuition in the selection of material.

I must now come to some methods of statistics. For illustrative material I shall use certain data supplied to me for the purpose by one of your members from his experiments on feeding pens of guinea pigs raw and variously processed milk in an effort to determine the influence of various methods of treating on the vitamin C content of milk. There were a great many trials involving 8 or 12 pigs each and 144 pigs in all. Let me take the data on boiled milk from one run. There were eight pigs. The first column gives their weight at the start, the second the number of days before scurvy developed. The numbers in the second column are added together and divided by the number (8) of pigs to find the mean length of time (18.75 days) before scurvy developed on the boiled milk ration. The third column gives the deviations of the individual numbers of days from this mean, and the fourth column the squares of these deviations. Means are found also for these two columns.

TABLE II

Weight at start	Days to scurvy	Deviation from mean	Deviation squared	Statistical constants
222	19	0.25	0.0625	$M = 18.75$
205	17	1.75	3.0625	$\delta = 0.875$
185	17	1.75	3.0625	$\sigma = 1.199$
185	19	0.25	0.0625	$p = 0.809$
190	19	0.25	0.0625	$E_m = 0.286$
370	21	2.25	5.0625	$E\sigma = 0.202$
335	19	0.25	0.0625	
195	19	0.25	0.0625	
Sum	150	7.00	11.5000	
Mean	18.75	0.875	1.4375	

The mean deviation is $\delta = 0.875$ days, the mean square deviation is $\sigma^2 = 1.4375$. The square root of the mean square deviation is called the *standard deviation*; in this case its value is $\sigma = 1.199$. The probable deviation has been defined geometrically on the (symmetric) frequency curve; it has also been defined for that case as that deviation than which the greater are equally numerous with the lesser deviations in the long run. As nothing is known relative to the ideal frequency distribution in this experiment other than the information given by the data themselves we must fall back on a formal arithmetic definition which is as follows: The probable deviation is $p = 0.6745\,\sigma$. This is merely the numerical relation between the probable deviation p and the standard error σ when the observations are infinitely numerous and distributed according to Gauss's law. If the true law of distribution of the data were not Gaussian the value $p = 0.6745\,\sigma$ might have small relation to the true probable error. For example, if the true law were Laplacean, the actual probable error would be not $0.6745\,\sigma$ but only $0.49\,\sigma$; the use of the arithmetic definition would give a probable error too large which

fortunately would be on the safe side (the figures in the first column of Table 1 would be 0.387, 0.150, 0.058, 0.022, 0.009, approximately, instead of the value given, but even these are far larger than the figures in column 2). The value of the individual observations may be written

$$\text{days} = 18.75 \pm 0.81; \quad \text{or} \quad 17.94 < \text{days} < 19.56.$$

The number of days it took to develop scurvy should lie half and half within the limits 17.94 and 19.56. The actual division is 5 within and 3 outside—pretty good. The largest deviation is 2.25 which is less than thrice the probable deviation, and although a deviation so large will occur according to the Gaussian law in the long run only once in 16 times (instead of 8), this need not disturb us.

Having calculated the standard deviation σ we may use the formulas applicable to the Gaussian distribution and write, if n be the number of observations (here 8),

$$E_M = \text{probable error in the mean} = 0.6745\,\frac{\sigma}{\sqrt{n}}$$

$$= 0.6745\,\frac{\sigma}{\sqrt{8}} = 0.286$$

$$E\sigma = \text{probable error in the standard deviation}$$

$$= 0.6745\,\frac{\sigma}{\sqrt{2n}} = 0.202$$

The probable error in the mean is the probable error in the individual observation divided by the square root of the number of observations. This is a definition because it does not apply to all frequency laws, and as we shall see has no relation whatsoever to the experimental results now under examination. In like manner the probable error of the standard deviation σ is the probable error ($0.6745\,\sigma$) of the individual observations divided by the square root of double the number of observations. These results should be written as

$$\text{the mean } M = 18.75 \pm 0.29,$$

$$\text{the standard deviation } \sigma = 1.20 \pm 0.20.$$

Now we have another set of data obtained in another trial on boiled milk. Omitting the details of the calculations, the results for the mean, the standard deviation and the probable error in the mean are:

$$M' = 13.75, \quad \sigma' = 6.58, \quad E'_M = 1.29$$

The difference of the means in the two runs is

$$M - M' = 5.00 = 17 \times (0.29).$$

The second trial gives a mean 17 times as far removed from the mean in the first trial as the probable error in that mean. Shall we say that this positively could not have happened? Or shall we say that the statistical method positively can not be applied here as per rule? Even if we use the much larger error of 1.29 figured from the data in the second case, the difference $M - M' = 5.00 = 4 \times 1.29$, nearly, and this can happen less than once in 20 times—according to the rule, which we don't believe (Table I, second column, Gauss's Law).

I have taken the data on the time to develop scurvy when the milk is boiled. There are three sets on raw milk, with these values:

$$M_1 = 43.00 \pm 2.78,$$

$$M_2 = 20.71 \pm 1.36,$$

$$M_3 = 27.00 \pm 2.18.$$

Taking the largest possible error of 2.78 we find that

$$M_1 - M_2 = 22.29 = 8 \times 2.78,$$

$$M_1 - M_3 = 14.00 = 5 \times 2.78.$$

According to the rule there are 7 chances in 10,000 for M_3 to be so far from M_1 and an almost incalculably small chance for M_2 to be so far away. Those who know the laws of the multiplication of chances can perhaps figure what is the chance by rule that both M_2 and M_3 shall be so far from M_1. What is the conclusion? I say beware of probable errors or rather of the mere formal application of them to meager statistical material. Your conclusions will almost certainly be wrong. The sound conclusion from the fact that the means of the different runs fail to lie reasonably close together is that the data are statistically inconsistent or insignificant. Permit me to refrain from defining the coefficient of variability —we have already too many definitions.

I know only one thing about feeding experiments —namely, that I eat what I like when I please and that my weight has remained between 157 and 167 pounds since before I left Yale in 1907. That is scanty foundation for participation in this symposium. I do not wish to give the impression that the data submitted to me for criticism are valueless; for all I know they may show just what it was hoped they would show; perhaps the experiments only meant this for a preliminary study wherewith to get his bearings—such studies are often necessary. As statistical material the individual runs, of which there are seventeen, show by their mutual comparison (as illustrated above) that for some reason they are inadequate statistically. Is it that 7 or 8 or even 12 pigs are too small a sample? If the animals in the time required to develop scurvy were as homogeneous as galvanometer deflections under good conditions these numbers would show no such inadequacy. As probable errors, according to the rule, vary inversely with the square root of the number of animals, must we take 800 pigs in a series, hoping thereby to reduce the errors to one tenth of their present values? "Pigs is pigs" would then be our only hope.

Looking over the data as a whole I have come to certain tentative statistical conclusions. Out of 144 pigs 3 starved to death sooner than drink milk, and all but 6 others, that died in brief periods, developed

scurvy in from 8 to 58 days on a diet of milk, whether processed or raw. I judge that guinea pigs were not made for a milk diet. If one could find laboratory animals which would live to a ripe old age on raw milk but go scurvy on treated milk he would, I should think, have a more desirable stock for this experiment. How about that very popular laboratory animal—Drosophila? Would the white rat do? One may be venturing too far afield in using guinea pigs. My choice would be Drosophila if possible. And why? Because the strains are pedigreed; they can be made to order, so to speak; the infinite variety of nature can be somewhat controlled in them. With a controlled pedigreed animal, mated brother to sister, and used in the F_1 generation only, one might hope to approach the narrow and reproducible experimental conditions of the physicist and chemist. In that recent and most interesting book which you have all just read, "The Biology of Death," by Raymond Pearl, you can even find a life table for Drosophila and thus make something of an allowance for natural death that might otherwise complicate lethal experiments on feeding. There is a cumulative value to Drosophila.

However, it may be possible to use guinea pigs despite their obvious aversion to milk. But in their use we should try to attain as great genetic or constitutional homogeneity as possible. First as to weights. The 144 pigs under consideration varied in weight from 420 to 132 grams. This is all too much variation. It may conceivably not be important what the weight is, and yet unless the indifference of weight has been clearly demonstrated we should restrict ourselves to pigs of approximately the same weight. Second, there is the matter of age. I have no idea what the variability of ages was nor whether it has been shown that age is a matter of indifference in its effect on the time required to develop scurvy. Still, it is safe to strive for homogeneity in the material, to reduce so far as may be the number of variables. Now as to weight I have done one very simple thing. I have taken all the pigs fed on raw milk except the one that would not drink it; there were 23 of them. I made a chart by plotting "days to scurvy" vertically and "weight at the start" horizontally—and lo! a strong correlation stared me in the face (Fig. 2).

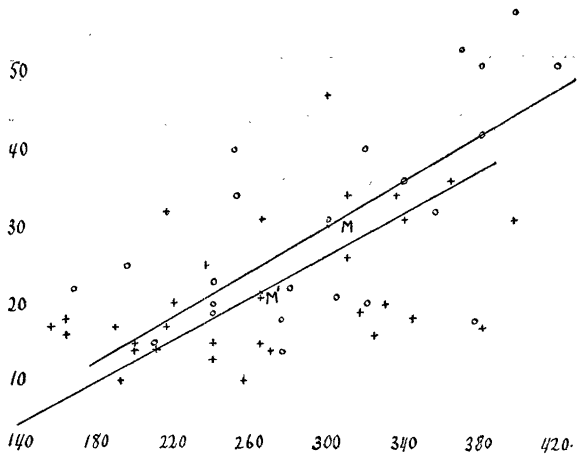

FIG. 2. A correlation diagram showing the tendency of the number of days (required to develop scurvy) to vary with the weight of the guinea pig at the start of the experiment. The abscissas (140–420) are the initial weights; the ordinates (10–50) are the number of days in which scurvy developed on a milk diet. Each circle represents these data for a single pig fed on raw milk and the upper line gives the general trend of these circles, M being the mean point. Each cross represents the data for a single pig fed on both pasteurized milk and the lower line gives their trend, M′ being the mean point. One may not compare the vertical distance between the means M and M′ without making allowance for the horizontal displacement between them.

It would take too long to define a correlation coefficient. In the "Biology of Death," pages 168–169, Pearl gives the definition in simple language for the intelligent οἱ πολλοῖ who frequent the Lowell Institute lectures. Suffice it here to state that the coefficient I found is $r = 0.68 \pm .07$. This is a high correlation and a small error. If I dared imitate Pearl on page 169 I should say that, the ratio of the probable error 0.07 to the coefficient 0.68 being over 9, the odds against such a correlation having arisen from chance alone are about 350,000,000 to 1; but I am too timid to follow him so far. It is enough for my purposes to note that the correlation is high and well established. I will not weary you with lines of regression or lines of closest fit, I have not calculated them. The diagram shows at a glance that for pigs fed raw milk, judging by this sample of 23, an addition of 100 grams in weight prolongs the time to develop scurvy by something like 15 days. This shows that the pigs should be uniform in weight or that the experimental data should be corrected for weight. It is not inconceivable to me that had I the ages of the pigs I should run into another correlation.

All the data for the 31 pigs fed bottle pasteurized milk were plotted up in like manner. The spread of points was again indicative of a sensible correlation coefficient, but I have not computed it nor estimated the correction for weight (except as one may do so roughly by glancing over the plotted points and guessing at a trend line). This should be done for each kind of treatment of the milk. One thing more I did. I figured the mean for raw milk and its probable error from the 23 available pigs and that for bottle pasteurized milk from the 31. The results were:

Raw $M_R = 30.65 \pm 2.4$ days. Avg. wt. $= 299$.
Bot. Past. $M_B = 21.4 \pm 1.0$ days. Avg. wt. $= 264$.

Without any correction for weight, the difference in the means, $M_R - M_B$, is 9.2, which is less than 4 times the probable error of M_R and as such would be just barely significant statistically when one considered the heterogeneity of the data and the great changes from sample to sample. An allowance for the difference of 35 grams in the average weights might easily reduce M_R to around 25 or 26 days and unless the probable error shrunk to less than 1.2 would eliminate whatever there was of statistical significance. If the probable error of M_B were used the result would be more hopeful, but in strict fairness we should use the formula for the probable error of the difference 9.2 of the two means as a function of their indiivdual probable errors; this would give a substantial indication of a significant statistical difference between the means provided, only no allowance for the weight correction be made.

We may raise the question of the comparison of the raw milk on the one hand and of the totality of treated milks on the other. I have not had time to compute the probable errors. The results for the means are:

				Reduced
Raw	M = 30.65 days	Avg. Wt. = 299		Raw
Bot. Past.	M = 21.4 "	" " = 264		26
Boiled	M = 18.7 "	" " = 238		22
Vat Past.	M = 18.2 "	" " = 252		23
Autoclaved	M = 17.4 "	" " = 232		22
Air Free	M = 15.5 "	" " = 237		22

The last column contains a rough estimate of the value of the mean for guinea pigs of the given average weights if fed raw milk. It is my judgment, though I can give no numerical estimate of the probability of that judgment, that the following conclusions are reasonably safe:

(1) On treated milk the pigs do develop scurvy sooner than on raw milk even when allowance for weight is made.

(2) The difference in the time required is smaller than I should have expected.

(3) There is no indication that the different ways of treating the milk produce statistically different results.

(4) An experiment simultaneously performed with sets of 25 pigs of like age, weight and sex and as homogeneous genetically as possible would probably give a good deal of significant statistical data. (The size of the litters from which the pigs were taken might have to be kept constant.)

In bringing this paper to a close I must plead the brevity of my time for preparation as an excuse for so inadequate a treatment of the large amount of data submitted to me and of similar data found in the literature on feeding experiments. Statistical work carefully done takes time—not merely time for routine calculations but far more time for thought. I am glad to know that statistical studies are arresting the attention of biochemists. The physicists and engineers of the Western Electric Company have found that they must resort to such methods when dealing with measurements of such inherently variable phenomena as the microphonic properties of carbon as used in telephone transmitters where the utmost care does not suffice to control the properties to the extent ordinarily attainable in physics. And as the use of the statistical method spreads, we must and shall appreciate the fact that it, like other methods, is not a substitute for but a humble aid to the formation of a scientific judgment. Only with this philosophy in mind may we truly hope, with care, to avoid, in the main, being classed in the superlative category of that oft-cited sequence of liars, damned liars and statisticians!

EDWIN BIDWELL WILSON
HARVARD SCHOOL OF PUBLIC HEALTH

THE SELECTION OF SUBJECTS FOR RESEARCH

THE question whether students should select subjects for research entirely of their own choice, or from a list of subjects proposed by their chosen professor, has been raised in many places and by numerous student generations, but I do not recall seeing any discussion of the subject. Wherever any considerable amount of research work is being done, it is important that the general policy be thoroughly understood in order that the *esprit de corps* may be maintained at the highest possible level.

Let us admit at the outset that almost any subject that one can suggest is worthy of investigation and that, *other things being equal*, the more lines that are being followed in a given laboratory the better. Diversity of interest has a broadening influence.

The trouble is that other things are never equal. No institution, no matter how large or how richly endowed, can possibly be equipped to do research work of an intensive character in more than a very few fields in which students may profess an interest. It is not too much to say that any institution which attempted to offer research facilities to meet the supposed needs of every student would descend to superficiality. It would receive for its pains the contempt of its graduates and the neglect of the public.

On the other hand, no institution is so small or poor that it can not do something to increase the sum of human knowledge, provided that it adheres unswervingly to a sufficiently narrow program, mapped out perhaps many years in advance of its possible realization. Such a program furnishes its own justification. Only one criterion must be met. Does the