AN EXPERIMENTAL STATISTICIAN LOOKS AT ANTHROPOMETRY

By Donald Mainland

Department of Medical Statistics, New York University College of Medicine, New York, N. Y.

The purpose of this paper is to record some of the reactions that occur when the content and methods of anthropometry are looked at by one who is now an experimental statistician in medicine, but was formerly an anatomist sufficiently concerned with human morphology to become a member of the American Association of Physical Anthropologists. The remarks can be somewhat systematized by examining each of the key words of the title.

Modern Anthropometry

Examining first the terminal word of the title, we consider the modern implications of "anthropometry" as revealed by one of its chief organs, *The American Journal of Physical Anthropology.* During the last five years, I have not been so intimately acquainted with the contents of this journal as I was during the preceding 20 years, but I have recently become immersed in it again, receiving both enlightenment and stimulation, for I find that its field of interest now includes such problems as taste testing, weight lifting, the cholesterol content of the blood, and the variation in urinary excretion patterns studied by chromatography. Doctor Stanley Garn[8] has shown what this expanding interest implies in the education of anthropologists, including training in experimental methods.

Statistics

Another key word in the title of this paper is "statistician." At a summer seminar held in 1951 by the Wenner-Gren Foundation for Anthropological Research, New York, N. Y., a very experienced and mature anthropologist remarked that a certain journal in his field was becoming "too statistical." There was no discussion, even to the extent of asking him which of the definitions of statistics (more than 100 in number[13]) he had in mind, or whether he was merely repelled by the mysterious formulae of mathematically sophisticated writers. If, however, his criticism went deeper than this, and indicated a suspicion that the mathematics masked doubtful assumptions, oversimplified the biology, and disregarded the poor quality of data—then, as an ex-anatomist and as an experimenter, I should often find myself in agreement with him.

But we cannot condemn an art simply because it is misapplied by some of its practitioners. The art that we are considering—experimenters' statistics—began in the first decade of this century, in a brewery chemist's laboratory in Dublin, Ireland, where W. S. Gosset, who published under the pseudonym "Student,"[24] devised a test, later called the t test, for evaluating the means of small samples of measurements. Thereafter, as is well known, the principles and techniques were extensively developed at the Rothamsted Agricultural

Experimental Station at Rothamsted, England, by R. A. Fisher[5, 6] (later Sir Ronald Fisher), and they are continuing to develop in many directions in the hands of many workers. They have spread so widely in applied science that there are very few goods, either grown or manufactured, that have not been influenced at some stage in their production by this form of statistics.

Basically, this variety of statistics expresses the logic of experimentation—the principles of inductive inference. The arithmetic is a subordinate part, but even if the arithmetic is complicated and laborious, it is easier than thoroughly good observation and experimentation, and very much easier than clear thinking. Also, the anwer that emerges from the arithmetic seems like the result of a mathematical proof. Hence the damage done by statistical "cookbooks" that have stressed the arithmetic and neglected the experiment. That introduces another word from the title—"experimental."

Experimentation

In so far as anthropometry becomes experimental, the anthropometrist will wish to use proper methods. To take a very simple problem, the comparison of two treatments of any kind, he will allocate the treatments to the subjects by what Fisher[6] called "a physical experimental process of randomization," such as a table of random numbers. Then, in interpreting his results, he will have only two possibilities to choose between—chance and the treatments—and he can easily find out how often chance, operating alone, would cause the difference that was produced in his experiment.

In seeking advice about experimental methods, however, the anthropometrist should not expect all professional experimenters still in circulation to be acquainted with modern methods, and some may scoff at randomization. Perhaps the best way to appreciate this procedure is to do an experiment or two in which it is properly used, and to note how "unclean" one feels in experiments where it has not been done. (Incidentally, all reports should indicate what the investigator means by "random," because the word is often still used very loosely. Haphazard sampling, or sampling without conscious selection, may be extremely nonrandom.)

The other great contribution of modern statistics to experimental method is the development of designs that not only avoid bias but enable one to obtain a maximum amount of information for a given expenditure of material, time, and money—that is, designs of high efficiency. Some of these designs are rather complicated but very valuable; for example, the recent use of a balanced, partially confounded design in the investigation of the various sources of error in photogrammetry of the human body.[11] In experimenting on human beings, it is usually desirable to keep to simple designs, for it is better to sacrifice some efficiency than to risk jeopardizing the whole experiment by breakdown or loss in a complicated scheme.[18] Randomized blocks, two-level factorials, and switchback designs are usually easy to apply. The switchback or reversal plan is very important where the sequence in which readings are taken may affect their magnitude, as in comparing right- and left-hand force by equipment that is unfamiliar to the subject.

Contrast of Surveys and Experiments

Although an anthropometrist may often experiment, most of the time he must use survey methods, observing what nature offers him, and yet acquaintance with experimental principles is essential for two reasons:

(1) Experimental statistics has had a profound influence on survey methods, as, for example, in the recent Canadian nation-wide survey[22] of statures, weights, and skinfold thicknesses.

(2) If we bear constantly in mind the outline of a simple experiment, we shall be acutely aware of the defects in even the best possible survey. A rather extreme illustration may be useful. An investigator wished to find out whether patients with rheumatoid arthritis differed from healthy persons in the concentration of a certain blood protein. He sent the figures to our department, suggesting a t test for comparison of the two means.

Naturally, we declined to apply any test at all, but we sent him a six-page memorandum telling him why we declined, and suggesting what he might do—briefly as follows:

(1) We recalled the essential difference between his survey and an experiment—no one can randomly allocate rheumatoid arthritis.

(2) We suggested that he prepare a list of factors, other than rheumatoid arthritis, that might conceivably cause a difference in protein levels between his arthritics and his controls (persons at a blood donor clinic). We began the list by mentioning 10 items: sex, age, racial stock, family relationships, diseases or disorders other than rheumatoid arthritis, treatment of patients, diet, physical activity, times and methods of blood collection, methods of chemical and other analysis.

(3) We suggested that he try to show which of these factors could probably be ruled out, either by internal evidence (from the data) or by external evidence (from other investigations).

(4) We pointed out that, after he had done all he could do, there would always remain some factors, either unsuspected or of unknown power, any one of which might cause bigger differences than chance would cause once in a thousand experiments—that, therefore, it would appear rather ludicrous to apply a significance test, as one would after an experiment.

Of course, anthropometrists can usually go much farther in removing bias from surveys than we can in medicine but, in the end, they also are left with the risk of bias from unknown factors, and herein lies the difference from experiments. In all experiments, there are unknown factors, but their effects have been randomly allocated, and, by our test of significance, we can insure ourselves against being "fooled" by these factors more than once in 20 experiments, or once in 100 experiments, whichever risk we choose to run.

Statistical Significance

One may well wonder whether significance tests should ever be applied to survey data, except to convince an investigator that a result that looks impressive could often occur by chance. That is, he would be asked to accept a verdict of nonsignificance ("not proved"), but to disregard a verdict of significance—a rather difficult psychological feat.

Many other things could be said about significance tests,[20] but only one will be mentioned here. To break the tyranny of the $P = 0.05$ standard, some anthropometrists have suggested that, with P between 0.05 and 0.01, the verdict should be "probably significant," but if P is less than 0.01, we should say "almost certainly significant." This may help some people, but it still seems to imply that a significance test can tell us what to think. All that it can do is tell us how often chance, operating alone, causes certain results. The decision then lies with us, and, whatever an investigator may say, or think he believes, the best way to see what level of significance he is adopting is to observe how he acts after he has found his P value.

Berkson's Fallacy: Competing Rates

Returning now to the biases that may lurk in surveys, we note one that is perhaps the most disturbing of all. In our department, we name it after Doctor Joseph Berkson[2] of the Mayo Clinic, Rochester, Minn., who demonstrated that, because hospital admission rates for different diseases differ, entirely erroneous conclusions can be reached regarding the relationship, or lack of relationship, between two diseases. Although this demonstration was published nearly nine years ago, it has only recently started to receive attention in medical statistics. Anthropologists may be better acquainted with it than their writings suggest. If they are not, the best way to start is probably by simplified arithmetic, as in the APPENDIX to this paper.

There appears, at present, no way of preventing or curing this bias, and what we should do in every survey, large or small, is ask: Could there have been present the two conditions, specified in the APPENDIX, that allow this competition between rates to come into play? To answer this question in the study of blood groups or prehistoric remains, we shall have to go back in thought many thousands of years, but we should consider carefully all the possible factors, and we should let our readers know that we have done so.

Anthropometry and Medicine

Berkson's fallacy has brought medicine prominently into the discussion, but the medical point of view obviously has influenced all the foregoing remarks. A short time ago Doctor Krogman[14] summarized the bearing of anthropology on medicine and dentistry and, from the medical side of the fence, as a member of a Study Group on Rheumatic Diseases, I hear, in our weekly discussions, clinicians and laboratory workers telling about their anthropometrical work—methods of assessment, morphological, physiological, and biochemical, in health and disease. It may, therefore, be worth while to consider a few examples of the way in which the three disciplines—anthropometry, medicine, and experimenters' statistics—could help each other.

(*1*) *Clinical normality.* This is a question that theoreticians have argued about for years, but physicians must act every day on some working definition. Here the anthropometrists could help, not only because they know the importance of standard techniques and the risk of errors in measurement, but because they are not so obsessed by disease as we are in medicine. They would be more likely to derive standards of normality from people performing their ordinary functions rather than in a hospital ward or under the so-called "basal" conditions of a metabolism room—a distinction that Doctor Gerda Seidelin[23] has rightly emphasized.

Assessment of normality is a statistical problem, and it is useful to recall the twofold classification of error now commonly used by statisticians in significance testing. Translated into terms of the assessment of a child's weight, a type 1 error occurs when we classify him as "underweight" when in reality his weight, although low in our standard range, is not low from pathological causes. A type 2 error occurs when we accept him as not underweight because it is, say, above the tenth percentile in our standard series, when in reality it is low from pathological causes. This classification of error is not merely convenient shorthand; it focuses our attention on the two risks that we run in all statistical judgments, and it helps us to set our standards. Particularly, it shows the need for knowledge of the range of variation found in each disease—"norms" for diseases.

(2) *Differential diagnosis.* The sexing of a skull, like the physician's attempt to decide which of two or more diseases a patient suffers from, raises the question of observation versus measurement—the same question as somatotyping versus anthropometry. The question can be settled only by thorough comparisons of the two methods, and often a combination of both will be found most valuable. When a metrical method involves the combination of measurements of several features on each subject, the appropriate way to evaluate the method is to get the best combination of measurements by discriminant functions. When Fisher[5] first described this technique in the 1938 edition of his *Statistical Methods,* he illustrated it by reference to the sexing of mandibles, but his technique has been curiously neglected in anthropometry until recently.[3]

(3) *Psychological problems.* As anthropometrists turn more from bones to living subjects, and from anatomy to physiology, they will meet more problems of psychology, familiar to clinical research workers who try to obtain healthy controls. In our own studies of bone density, for example, it was not difficult to pick up in a few minutes a dozen or more students who were willing to be X-rayed, but several hours spent on personal appeals to business men would sometimes bring in only one or two. It may not be very difficult to induce Australian aboriginals of both sexes and all ages to be photographed in the nude, but to obtain many middle-aged and old citizens of New York for somatotyping would not be so easy. Such a restriction can lead to distorted conceptions, as it has done in biochemical tests for the detection of cancer.[12] Many of these tests have seemed successful at first because controls have been interns, residents, nurses, and young laboratory workers, whereas the cancer patients have been older people, and the tests have been simply detecting age differences.

It must be remembered, also, that those who volunteer as subjects for research are likely to differ from those who do not volunteer—psychologically,[15] therefore presumably also in their physiology, biochemistry, and, not improbably, in their morphology.

(4) *Deceptiveness of familiar statistics.* Significance tests are now common in anthropometry and medicine, but in both fields a question that is often of

greater importance is: How large or how small may be the true (population) value of this thing that I have tested? That is, what are the confidence limits? These limits often reveal how little we know about the sampled population, whatever the significance test may have shown.[16, 19]

The coefficient of correlation has played a large part in anthropometry during and since the time of Karl Pearson, who contributed so much to biometry in the pre-experimental period. The correlation coefficient is retained in experimental statistics chiefly because of its arithmetical relationship to the linear regression coefficient. Thus, its probability P, which can be found from tables,[7] is exactly the same as the P for the corresponding regression coefficient, and this saves the trouble of a special test. Also, the correlation coefficient (r) provides (by the use of $1 - r^2$) a quick method of finding the residual variation —the scatter of dots about the regression line. The full implications of correlation are, however, quite inappropriate to experiments, because the values of the independent variate are selected, and the same may be true of surveys. Besides the inappropriate use of r, three features are regrettably still common in anthropometric and medical literature:

(a) Standard errors affixed to correlation coefficients, although Fisher[5] pointed out 30 years ago how dangerous this could be.

(b) The use of complex variables, such as ratios and indices, in correlation and regression without previous consideration of what the effects may be— a point to which Gavan[10] has called attention. Sometimes the situation is made even more complicated by the use of the same measurement in forming the X and Y variates.

(c) The use of the correlation coefficient as adequate evidence of the reliability of a technique—to show, for example, how closely two observers agree, or how closely one relatively easy method of measurement agrees with a standard but more difficult method. If the correlation coefficient is used at all, the next step should be to estimate the residual variation. For example, if the coefficient of correlation is +0.995, all except 1 per cent of the total variation between readings by the easier method would be registered also by the standard method, but this 1 per cent, expressed as actual units of measurement, may show that the easier method is by no means precise enough for the purpose in hand. Instead of correlation or regression, however, a familiar simple technique is usually preferable. From the series of differences (reading by standard method *minus* reading by easier method) find the mean difference, *i.e.*, the systematic difference between the two methods, and find the standard deviation of the series, *i.e.*, the variable or random difference between the methods. (When the relationship between the readings by the two methods is curvilinear, special techniques of estimation, of course, have to be employed.)

(5) *The need for further exploration*. The foregoing remarks are not meant to imply that experimenters' statistics can answer all the questions that anthropometrists may ask. It is still a young art, and bears the marks of its origin in biology, where variation between individuals, after elimination of detectable major causes of difference, is often not very far from normal (Gaussian) in form, and where, as sampling experiments have shown, even considerable skewness does not vitiate the t test or analysis of variance. But the techniques are now

being used outside biology. In the work in our own laboratory, for example, we need to know how far they are safe when applied to densitometer readings on X-ray films. We find that our interest is turning very much to techniques that do not assume any form of distribution, Gaussian or other (nonparametric techniques).

Experimenters' statistics retains, also, its interest in averages and in over-all variation, but in assessing clinical normality we are interested in the tails of skew and possibly irregular and truncated distributions. Healthy subjects, as was mentioned above, are not plentiful, and therefore we might think of pooling data from different centers. But even the best laboratories, employing the same methods, differ in their results. Therefore, to pool data might make the range too wide to be of use in any of the centers. Hence our question is: What can be done with "local norms" derived from relatively small samples? Only experiments can tell us.*

In many other directions also, such as factor analysis, exploration is necessary, and this youthful art of experimenters' statistics needs guidance and constructive criticism by those who become informed of what it has done and is trying to do. It could be helped greatly by forward-looking anthropometrists.

Appendix

Berkson's Fallacy: Competing Rates

The original discovery. An early demonstration of Berkson's[2] discovery occurred in a search for a possible association between diabetes and cholecystitis. There was such a strong impression of the existence of this association that some surgeons were removing gallbladders in the treatment of diabetes. To test the soundness of this belief, the incidence of cholecystitis in diabetic patients was compared with its incidence in persons who came to the clinic for eye testing, because it could not reasonably be suspected that there was any association between cholecystitis and errors of refraction. The frequency of cholecystitis was found to be higher in the diabetics by an amount that was statistically significant. Then Berkson showed that such results could be entirely fallacious.

Extension to other surveys. The fallacy can affect any kind of survey of any material, living or dead, organic or inorganic. In demonstrating it we can use the symbols A, B, and X to denote any feature, quality, or attribute, and we can retain the hospital term "admission rate." The admission rate for any attribute is the proportion of the possessors of that attribute that appear in a collection. Thus, in many anatomy departments the admission rate for males (A's) is higher than for females (B's). Also, there is often a tendency to admit specially, or to preserve when found, specimens with peculiar features (X's). Substituting a simple arithmetical example for Berkson's general proof, we proceed as follows.

Simple arithmetic. Three attributes: A, B, and X. In the population: 1000 persons with A, 1000 with B. In each of these, 100 persons have X also. Therefore there is no closer association between B and X than there is

* In this exploration, the New York University Department of Medical Statistics, New York, N. Y., is fortunate in having financial aid from Eli Lilly and Company, Indianapolis, Ind., and the encouragement of the Company's Counselor on Research Grants, Doctor Donald D. Van Slyke.

between *A* and *X*. The admission rates to a certain survey are: for *A*, 50 per cent; for *B*, 20 per cent; for *X*, 40 per cent. To find how many persons with each attribute will come into the survey:

Group *A*, *X*. Total persons = 100. Fifty per cent of these, *i.e.*, 50 persons, are admitted because they have attribute *A*, leaving 50 outside. Of these latter, 40 per cent (20 persons) will be admitted because they have attribute *X*. Total admissions = 70. (The same result is obtained if 40 per cent of 100 are admitted first because they have attribute *X*, and then, from the remaining 60, 50 per cent are admitted because they have attribute *A*.)

Group *A*, *not-X*. Total persons = 900, of whom 450 are admitted because they have attribute *A*.

Group *B*, *X*. Total persons = 100, of whom 20 per cent (20 persons) are admitted because they have attribute *B*, leaving 80 outside. Of these latter, 40 per cent (32 persons) are admitted because they have attribute *X*. Total admissions = 52.

Group *B*, *not-X*. Total persons = 900, of whom 180 are admitted because they have attribute *B*.

In summary, the following persons will be found:

	X	*not-X*	*Total*
A	70	450	520
B	52	180	232
Total	122	630	752

The percentage frequencies of *X* are as follows: Of the *A's*, $70 \times 100/520 =$ 13.46 per cent have attribute *X*. Of the *B's*, $52 \times 100/232 = 22.41$ per cent have attribute *X*.

The difference (8.95 per cent) is very significant, for chance would cause such a difference less than once in 300 experiments. Something more than chance was operating, but it was not the closer association between *X* and *B* than between *X* and *A*. It was the difference in admission rates—a kind of competition between rates. The rate in *A* offers stronger opposition to *X* than does the rate in *B*; *i.e.*, more of the *B's* are left to be admitted because of *X*.

When division is more than dichotomous (*A*, *B*; *X*, *not-X*) the effects are more difficult to visualize, but they must introduce bias in the same way.

Note. Instead of attributes, the symbols can represent measurements, *e.g.*, age in ascending order (*A*, *B*, *C*, *etc.*) and weight (*X*, *Y*, *Z*, *etc.*).

Necessary conditions. Using general terms, we can state the two conditions necessary for the occurrence of the bias:

(1) The admission rates of *A* and *B* must be different.

(2) The admission rates of *X* and *not-X* must be different.

The masking of a real difference. The bias may mask a real difference, as the reader can demonstrate by doing the arithmetic of the following example. Let all admission rates and the calculation for *A's* remain as in the above example, but let 60 of the *B's* have *X*, instead of 100 (*i.e.*, 6 per cent instead of 10 per cent). (When, in the calculation, 40 per cent of 48 is found to be 19.2, call it 19 persons.) Answer: 14.16 per cent of the *B's* in the survey are *X's*.

Further information. Further discussion of this fallacy can be found in articles by Berkson[2] (the original demonstration), Mainland[17] (application to autopsy data), and Mainland and Herrera[21] (application to clinical surveys).*

Suggestions in the search for Berkson's fallacy. It is instructive to consider how the bias might arise in various fields, and the five speculations initiated below will serve their purpose if they prompt specialists in each field to replace them by more realistic examples.

(1) Selective admission of unusual specimens into anatomical collections has been mentioned already, but even in departments that preserve the bones of all their cadavers, numerous selection factors have operated before the cadaver arrives—social factors, physique differences, economic status, racial differences (even within the White group). Secular changes occur in the strength of all these factors, including differences in the type of subjects during a severe economic depression. In the investigation of skeletal stature (and the comparison with living subjects' stature) careful thought should be given to the bias due to competing rates.

(2) In the investigation of prehistoric remains, it should be remembered that the preservation of the subjects, and therefore their availability for study, depends partly upon the terrain, and two types, A and B, may have lived in different terrains. To qualify as the attribute X, this morphological feature would have to be associated with something (perhaps sex, age, or occupation) that tended to promote preservation, whichever terrain its possessors lived in.

(3) The text of the article has referred to psychological differences between those who volunteer as subjects for research and those who do not. We may suppose an association between this and some morphological or physiological feature that is to be investigated (X), and if there is also a difference, independent of X, in volunteer rate between sexes or racial groups (A, B) or ages (A, B, C, *etc.*) the stage is set for Berkson's fallacy.

(4) In the examination of soldiers at a separation center for the study of differences in somatotype associated with age, two possibilities would be relevant:

(a) Differences in admission rate to the army associated with age (A, B, C, *etc.*).

(b) Differences in admission rate (or survival rate) associated with physique. Some evidence of association between somatotype and occupation[9] should be recalled in this connection.

(5) In blood group research, if the groups exert selective effects (or are associated with selection factors)[4] or if there is an association between groups and disease,[1] the problem of bias appears to fall into the basic pattern described by Berkson.

References

1. AIRD, I., H. H. BENTALL, J. A. MEHIGAN, & J. A. FRASER ROBERTS. 1954. The blood groups in relation to peptic ulceration and carcinoma of the colon, rectum, breast and bronchus. Brit. Med. J. **2:** 315.
2. BERKSON, J. 1946. Limitations of the application of fourfold table analysis to hospital data. Biometrics Bull. **2:** 47.

* There has recently appeared an important article by Doctor Berkson (Proc. Mayo Clinic, 1955. **30:** 319) that recounts the history of his discovery before its application to the cholecystitis problem and shows how the bias can produce spurious associations in longitudinal forward-going studies, such as the search for an association between smoking and lung cancer.

3. Bronowski, J. & W. M. Long. 1952. Statistics of discrimination in anthropology. Am. J. Phys. Anthrop. **10:** 385.
4. Brues, A. M. 1954. Selection and polymorphism in the A-B-O blood groups. Am. J. Phys. Anthrop. **12:** 559.
5. Fisher, R. A. 1950. Statistical Methods for Research Workers. Hafner. New York, N. Y.
6. Fisher, R. A. 1951. The Design of Experiments. Hafner. New York, N. Y.
7. Fisher, R. A. & F. Yates. 1953. Statistical Tables for Biological, Agricultural and Medical Research. Hafner. New York, N. Y.
8. Garn, S. M. 1954. On the education of the physical anthropologist. Am. J. Phys. Anthrop. **12:** 607.
9. Garn, S. M. & M. M. Gertler. 1950. An association between type of work and physique in an industrial group. Am. J. Phys. Anthrop. **8:** 387.
10. Gavan, J. A. 1952. "Growth of Guamanian children"—some methodological questions. Am. J. Phys. Anthrop. **10:** 132.
11. Geoghegan, B. 1953. The determination of body measurements, surface area and body volume by photography. Am. J. Phys. Anthrop. **11:** 97.
12. Hammond, E. C. Personal communication.
13. Kendall, M. G. 1947. The Advanced Theory of Statistics. **1:** 1. Griffin. London, England.
14. Krogman, W. M. 1951. The role of physical anthropology in dental and medical research. Am. J. Phys. Anthrop. **9:** 211.
15. Lasagna, L. & J. M. von Felsinger. 1954. The volunteer subject in research. Science. **120:** 359.
16. Mainland, D. 1952. Elementary Medical Statistics: The Principles of Quantitative Medicine. Saunders. Philadelphia, Pa.
17. Mainland, D. 1953. The risk of fallacious conclusions from autopsy data on the incidence of diseases, with application to heart disease. Am. Heart J. **45:** 644.
18. Mainland, D. 1954. The planning of investigations. *In* Methods in Medical Research. **6:** 138. Year Book Publ. Chicago, Ill.
19. Mainland, D. 1954. Confidence limits. *In* Methods in Medical Research. **6:** 191. Year Book Publ. Chicago, Ill.
20. Mainland, D. 1954. Standards of significance. *In* Methods in Medical Research. **6:** 195. Year Book Publ. Chicago Ill.
21. Mainland, D. & L. Herrera. 1954. Clinical surveys. *In* Methods in Medical Research. **6:** 159. Year Book Publ. Chicago, Ill.
22. Pett, L. B. 1955. A Canadian table of average weights. Can. Med. Assoc. J. **72:** 12.
23. Seidelin, G. 1952. The concept of normal. Acta Med. Scand Suppl. 266. **142:** 891.
24. "Student." (W. S. Gosset.) 1908. The probable error of a mean. Biometrika. **6:** 1.