STATISTICS FOR THE INVESTIGATION OF INDIVIDUAL CASES*

R. W. PAYNE AND H. GWYNNE JONES

Institute of Psychiatry, University of London, Maudsley Hospital

PROBLEM

Much of the work of a clinical psychologist consists of making relatively routine psychological measurements of fairly well established traits, either cognitive or orectic. It is well known, however, that there can be no measurement without error. The psychologist must have the means of taking error into account if he is to assess his test scores intelligently. There appear to be three main types of question which face clinical psychologists:

1. *The Abnormality of a Discrepancy between Two Scores*

This problem arises every time a psychologist gives more than one measure. Perhaps the commonest example is the Wechsler-Bellevue Intelligence Scale. This test provides two rather different measures of intelligence, the "Verbal Scale IQ" and the "Performance Scale IQ". It is a common experience that these two scores are divergent. In fact the discrepancy may suggest interesting hypotheses in line with other abnormalities the patient shows. However, before we can assess such a discrepancy, we must take into account two factors. We know that neither scale is perfectly reliable and we also know that the scales are not perfectly correlated. Therefore, many normal people would show discrepancies between the two scales which one need not take seriously. The first question we can ask ourselves then, is how frequently would a discrepancy as large as the one we observe occur in the normal population? That is, how "abnormal" is the difference we observe between our test scores?

2. *The reliability of a discrepancy between two scores*

In certain cases, we may have occasion to give two tests which measure rather different traits. For example, we may give a test of long term retention, and a test of general intelligence. It may be the case that these tests have a very low intercorrelation in the general population, so that quite large discrepancies between these scores could be quite "normal" or usual in the general population. Nevertheless on clinical grounds, we might expect our patient to have a lower memory test score than a general intelligence test score. We are not implying that this would be an abnormally *large* discrepancy. Many people may have as large differences. We *are* implying, however, that it is a *measurable* difference. We know that neither test is perfectly reliable, so that small differences will occur by "chance". What we wish to know is how large a difference between any two scores must be before we can be sure the difference could not be due merely to error of measurement of the tests.

3. *Testing a Clinical Prediction*

A third type of problem is slightly different. Very often the clinical psychologist finds himself repeating a measurement with a certain expectation or "prediction". For example, a patient may obtain an "average" IQ when first seen. Two years later, there may be strong clinical grounds for believing that deterioration has taken place. We, therefore, wish to retest him on the same (or a similar) test of intelligence to confirm the hypothesis that he has deteriorated. We may, indeed, find that his score is now below average. Have we in fact confirmed our hypothesis?

Again we know that tests are not perfectly reliable and that such changes in score occur in perfectly normal people. Essentially we need a control group. We need to know what proportion of individuals like our patient, of the same IQ on the first

*Editorial Note: The authors use the term "standard error of prediction" which is customarily called a "standard error estimate" in America, and in this country we uniformly refer to the "proportions" or "probabilities" corresponding to various standard score values instead of the term "percentile".

test, and who have *not* deteriorated would show an equal drop in IQ on retest. If the figure is fairly large, of course, our result does not prove that deterioration has really occurred. The practising psychologist will not have time to conduct the appropriate control experiment. Is there any other way of providing an approximate answer?

Suggested Solutions

The clinical teaching section of the Institute of Psychiatry (Maudsley Hospital) became aware of these problems several years ago. The following simple statistical models were eventually evolved and have since been applied routinely in clinical practice.[1]

1. *The Abnormality of a Discrepancy Between Two Scores*

Let us call the two raw scores X and Y. We are required to discover how frequently a discrepancy or difference between them occurs in the general population. Let us call this difference D, which is merely X - Y. Provided that the bi-variate distribution of the two scores is normal, the percentile value for any D score can be obtained from the ordinary table of the normal curve relating percentiles to standard scores. All we need to do, is to express our D score as a standard score in the usual way, substituting in the formula:

$$Z_d \text{ (standard "D" score)} = \frac{D - \bar{D}}{\sigma_d}$$

In this equation $D = X - Y$, $\bar{D}$ (mean D) $= \bar{X} - \bar{Y}$, and σ_d (standard error of difference, or standard deviation of the D. scores) $=$

$$\sqrt{\sigma_x^2 + \sigma_y^2 - 2r_{xy}\,\sigma_x\,\sigma_y}$$

Where σ_x is of course the standard deviation of test x, σ_y the standard deviation of test y, and r_{xy} the Pearson product-moment correlation between tests x and y.

This general method will tell us how frequently this particular D score occurs in the standardization population. Our D score, however, was the difference between two raw scores, and as such is influenced by both X and Y. If, however, test x has a very much larger standard deviation (and range) than test y, score D will be influenced much more by score X than score Y (*i.e.*, D scores will correlate higher with X than with Y scores). In an extreme case, an "abnormally" (infrequently) large X score would automatically produce an "abnormally" large D score. This would be the case if raw X and Y scores are used. Psychologically, however, we are not really concerned with the discrepancy between raw X and Y scores, as raw scores on most psychological tests are quite arbitrary. What we are concerned with is a discrepancy between the *percentile* standing on test x and the *percentile* standing on test y. Thus, it is really the "abnormality" (frequency) of a discrepancy between two percentiles that concerns us. We can easily estimate this if we first express our raw X and Y scores as *standard* scores, and then assess the frequency of the difference between these two standard scores in the general population.

Thus, we can first transform our raw X and Y scores into standard scores according to the formulae:

$$Z_x = \frac{X - \bar{X}}{\sigma_x} \text{ and } Z_y = \frac{Y - \bar{Y}}{\sigma_y}$$

We can then find the difference between the two standard scores: $D_z = Z_x - Z_y$ (D_z = difference between standard scores). This difference (discrepancy) can then

[1]The use of the regression equation to answer the third type of problem has been illustrated by Slater[4] in a slightly different form. The techniques to be discussed were largely initiated by Dr. A. Lubin, Dr. M. B. Shapiro and Professor H. J. Eysenck in mutual discussion. The formulae quoted in subsequent sections for the standard error of a difference, and the regression equation are to be found in any standard textbook of statistics, for example P. O. Johnson[2] or Q. McNemar[3].

itself be expressed as a standard score, so that its percentile position can be ascertained:

$$Z_{d_z} = \frac{D_z - \bar{D}_z}{\sqrt{\sigma_{z_x}^2 + \sigma_{z_y}^2 - 2r_{xy}\,\sigma_{z_x}\,\sigma_{z_y}}}$$

However, $\bar{D}_z$ (the *mean* difference between standard scores of x and y) is zero, as standard scores are made to have a mean of zero. Similarly σ_{z_x} (the standard deviation of Z_x. or x standard scores), is one, as standard scores are made to have a standard deviation of one. Thus:

$$Z_{d_z} = \frac{Z_x - Z_y}{\sqrt{1 + 1 - 2r_{xy}}}$$

By consulting the tables for the normal curve, we can transform this Z into a percentile and discover how frequently the difference occurs in the general popular tion. A "two tailed" test tells us how frequently a discrepancy of this magnitude o- greater between the two standard scores occurs in the general population, regardless of the direction of the discrepancy. A "one tailed" test tells us how frequently the difference occurs in this particular direction (for example how frequently Z_x exceeds Z_y by this degree in the population).

Example:

A man of 25 obtains a Wechsler[6] Verbal Scale IQ of 120, and a Wechsler Performance Scale IQ of 105. We wish to know how common such a large discrepancy actually is. We know that the mean of both Wechsler scales is 100, and the standard deviation is 15. Thus, the Verbal Score, expressed as a standard score $= \frac{120 - 100}{15}$ $= 1.33$. The Performance IQ expressed as a standard score is $\frac{105 - 100}{15} = 0.33$.

The difference between these two standard scores (Verbal minus Performance) is 1.00. The correlation between the Wechsler Verbal IQ and Performance Scale IQ is 0.71(6 p. 124)[2]. Thus the standard error of the difference between these two Wechsler standard scores would be:

$$\sqrt{1 + 1 - 2(0.71)} = 0.762$$

Dividing the difference (minus the mean difference – in this case zero) by the standard error of the difference, we find that:

$$Z = \frac{1.00}{0.762} = 1.31$$

Consulting the table for converting standard scores into percentiles(2, p. 359) we see that this corresponds to a percentile value of 90.5. Thus we see that 19% of the standardization population would have a discrepancy this large (or larger) between Verbal and Performance IQ (the "two tailed" test); 9.5% of the standardization population would have a discrepancy this large (or larger) in favor of the Verbal IQ, as in this case; and 9.5% of the standardization population would have a discrepancy this large (or larger) in favor of the Performance IQ (the "one tailed" test). Whether this discrepancy is infrequent enough in the general population to be regarded as "abnormal", and perhaps worthy of further investigation, must be left to the judgment of the individual clinical psychologist.

---

[2]We would not use the value of .83 which Wechsler also quotes after correction for attenuation, as we are concerned here with the empirically found range of verbal-performance discrepancies in the standardization population.

2. *The Reliability of a Discrepancy Between Two Scores*

Let us again call the two raw scores X and Y. We are required to discover whether the discrepancy between them (D, or X - Y) is large enough to be outside the range of differences attributable to the errors of mesaurement of the two tests. The "standard error of measurement" is the usual psychological estimate of the range of error which we can expect from a single test. The standard error of measurement is usually defined according to the formula:

$$S.E._m = \sigma_x \sqrt{1 - r_{xx}}$$

where r_{xx} refers to the coefficient of reliability of the test concerned. This is either a test-retest product moment correlation coefficient, a "split-half" product moment correlation coefficient corrected for attenuation, or sometimes a form *vs.* form correlation coefficient.

If we accept the equation for the standard error of measurement, then the probability of an obtained score "X" on any test departing from a given score which we could call "T" through error of measurement, can easily be obtained according to the formula:

$$Z = \frac{X - T}{\sigma_x \sqrt{1 - r_{xx}}}$$

Thus $(\sigma_x \sqrt{1 - r_{xx}})^2$ represents the error variance of test x. However, we know that the standard error of a *difference* between two scores X and Y is given by the formula:

$$\text{SE diff} = \sqrt{\sigma_x^2 + \sigma_y^2 - 2r\,\sigma_x\,\sigma_y}$$

If we replace the variance of X by the error variance of X and the variance of Y by the error variance of Y in the above equation, we will estimate the standard error of the differences attributable solely to error. As error is uncorrelated, however, the last term $(-2r\,\sigma_x\,\sigma_y)$ will become zero. (Errors of measurement on test x are uncorrelated with errors on test y). Thus the standard deviation of the discrepancies between X and Y attributable solely to *error of measurement* as defined, will be given by the formula:

$$\sigma \text{ error diff} = \sqrt{(\sigma_x \sqrt{1 - r_{xx}})^2 + (\sigma_y \sqrt{1 - r_{yy}})^2}$$

If we wish to determine the probability of any difference (X − Y) between two scores being due solely to the errors of measurement of the two tests, we can thus set up the following ratio:

$$Z = \frac{D - \bar{D}}{\sqrt{(\sigma_x \sqrt{1 - r_{xx}})^2 + (\sigma_y \sqrt{1 - r_{yy}})^2}}$$

where:

$$D = X - Y \text{ and } \bar{D} = \bar{X} - \bar{Y}$$

As we have seen, however, D, the difference between X and Y in raw scores, can be correlated much more highly with X than with Y, if σ_x is much greater than σ_y. In an extreme case for example, where the range of X is very much larger than the range of Y, an extremely high X score would always be "reliably" different from a Y score of *any* value, whatever the reliability of test Y, if the above formula were used. In other words, the judgment of whether or not the difference D were large enough to be measurable, would be almost entirely a function of the size of X. This is clearly not what is required. Again, this difficulty can be surmounted if we first equalize the variance of X and Y by converting both scores to standard scores. We can then apply the formula above, and obtain a valid answer. In standard score units, the formula becomes:

$$Z = \frac{D_z - \bar{D}_z}{\sqrt{(\sqrt{1 - r_{xx}})^2 + (\sqrt{1 - r_{yy}})^2}}$$

$$= \frac{D_z}{\sqrt{(1 - r_{xx}) + (1 - r_{yy})}}$$

where:

$D_z = Z_x - Z_y$, $\bar{D}_z = \bar{Z}_x - \bar{Z}_y = 0$,
r_{xx} = coefficient of reliability of test x, and
r_{yy} = coefficient of reliability of test y.

Our standard score, Z, can be converted into a percentile by consulting the table for the normal curve in the usual way.

Example:

A male patient of 20 was given the "General Aptitude Test Battery" of the United States Employment Service.[5] It was expected from his Wechsler results that his Spatial Aptitude would be higher than his "G" score on the GATB. He obtained a "G" score of 75 and a "Spatial" score of 86. We wish to discover whether or not this difference is in fact measurable, that is, whether a difference this large is within the range of differences one might expect from error of measurement.

If we call the "Spatial" score "X" and the "G" score "Y", we can first convert these scores into standard scores. We know that for both these tests, the mean is 100 and the standard deviation is 20. Thus:

$$Z_x = \frac{X - \bar{X}}{\sigma_x} = \frac{86 - 100}{20} = -0.70 \text{ and } Z_y = \frac{Y - \bar{Y}}{\sigma_y} = \frac{75 - 100}{20} = -1.25$$

The coefficient of reliability of the Spatial test (test "X") is .88, and the coefficient of the "G" test (test "Y") is .94. These are test-retest reliability coefficients quoted in the test manual[6, p. 1-4] for a male population comparable to the patient. We can now substitute in the equation:

$$Z = \frac{D_z}{\sqrt{(1 - r_{xx}) + (1 - r_{yy})}} \qquad \text{where } D_z = Z_x - Z_y$$

$$= \frac{0.55}{\sqrt{(1 - .88) + (1 - .94)}} = \frac{0.55}{0.424} = 1.30$$

Consulting the table for the normal curve we find that a Z of 1.30 corresponds to a percentile value of 90.3. Thus differences as large or larger than our eleven point discrepancy between "G" and "Spatial" scores could occur 19.4% of the time through error of measurement alone. Differences in the expected direction (*i.e.*, "Spatial" higher than "G") could occur 9.7% of the time through error of measurement (the one tail test). Thus we cannot be certain that this test discrepancy represents a "real" difference between these two aptitudes, and the assessment of these probabilities must again be left to the individual clinician.

3. *Testing a Clinical Prediction*

If we have given a certain test to a patient, and expect a drop in score following some trauma or other circumstance, and subsequently retest our patient on the same test, we wish to discover what proportion of subjects, all of whom had the same *initial* score on the test, but who did *not* suffer the trauma, would be expected to show a drop this large. This is a relatively simple matter, if we have the relevant data for the test in question. What is required is the test-retest data for the test over an equivalent period of time, for a representative sample of subjects similar to the patient. In other words, our patient's original test score must fall within the distribution of scores obtained by our statisical "control group" on the initial testing.

Let us call the initial test score X, and the score obtained on retest, Y. The regression line of Y upon X is that straight line which best fits (by "least squares") the points represented by the *mean* Y score of each group (column) of subjects with homogeneous X scores. If we use $\hat{Y}$ to represent that Y score which would be predicted from the regression equation from our patient's X (initial test) score, then $\hat{Y}$ also represents the *mean* Y score (retest score) of all those subjects who started out with the same X (initial test) score as our patient. It is this group then, which serves as our patient's "control" group. The *standard error of prediction* using this regression equation represents the standard deviation of the Y (retest) scores of all those patients who started out with the same X (test) score as our patient. We have only to determine then, whether our patient's Y (retest) score falls within or outside the distribution of Y (retest) scores obtained by those people with the same initial X (test) score.

In practice then, we would first discover the *average* Y (retest) score of those people with the same initial X (test) score as our patient, by substituting in the regression equation:

$$\hat{Y} = a + b\,X \quad \text{where} \quad a = \bar{Y} - b\,\bar{X} \qquad \text{and } b = r_{xy}\frac{\sigma_y}{\sigma_x}$$

$\hat{Y}$ = mean retest (Y) score of people who start out with score X
X = patient's initial score
$\bar{X}$ = Mean of all subjects on test X (test)
$\bar{Y}$ = Mean of all subjects on test Y (retest)·
σ_x = standard deviation of all subjects on test X (test)
σ_y = standard deviation of all subjects on test Y (retest)
r_{xy} = product moment correlation between test x and test y (i.e. test-retest correlation)

Having discovered our value for $\hat{Y}$, we can then discover the *standard deviation* of the Y scores (retest scores) of those people with the same initial X (test) score. This is the standard error of prediction, and is merely:

$$\text{S.E.}_{\text{predic.}} = \sigma_y \sqrt{1 - r_{xy}^2}$$

Having obtained this value, we merely determine whether our patient's score falls within or outside this group of retest scores for people with the same initial score, by expressing our subject's score as a standard score for this distribution according to the formula:

$$Z = \frac{Y - \hat{Y}}{\text{S.E.}_{\text{predic.}}}$$

The final Z value can be converted to a percentile by consulting the standard table of the normal curve. Since a prediction has been made in each case where this statistical model is used, a *one tailed* test of significance is appropriate.

Example:

A 30 year old schizophrenic patient is given the Wechsler Bellevue Form I, and obtains a total weighted score of 110 (IQ = 112). He is believed to have deteriorated intellectually following treatment with Serpasil, and is retested one month later. He now obtains a total weighted score of 83 (IQ = 94) on Wechsler Form I. We wish to discover how many schizophrenics of this initial IQ who have *not* had any treatment intervening between test and retest on the Wechsler over a four week period, would have dropped to this extent.

Some control data is given in an article by Hamister[1] who tested a group of 34 schizophrenics on Wechsler I and retested them four weeks later. Their mean total weighted score on the first test was 92.97, with a S.D. of 26.00. Their mean total weighted score on retest was 104.26, with a S.D. of 28.91. The correlation between test and retest was .84.

First we must use the regression equation to discover what the average retest score ($\hat{Y}$ in the formula) was in Hamister's group, for patients with the same initial Wechsler score:

$$\hat{Y} = a + b\,X$$

$$\text{where } a = \bar{Y} - b\,\bar{X}$$

$$\text{and } b = r_{xy}\,\frac{\sigma_y}{\sigma_x}$$

We know that $\bar{Y} = 104.26$ (retest mean, $\bar{X} = 92.97$, $\sigma_y = 28.91$, $\sigma_x = 26.00$, $r_{xy} = .84$, and X (patient's initial score) = 110, and Y (patient's retest score) = 83

Therefore:

$$b = .84 \times \frac{28.91}{26.00} = 0.93$$

and:

$$a = (104.26 - 0.93) \times 92.97 = 17.80$$

Thus:

$$\hat{Y} = (17.80 + 0.93) \times 110 = 120.10$$

The standard deviation of retest (Y) scores of schizophrenics with this initial X score is given by the formula:

$$\text{S.E.}_{\text{predic.}} = \sigma_y \sqrt{1 - r_{xy}^2} = 28.91 \sqrt{1 - 84^2} = 15.69$$

Thus the patient's standard score position in the distribution of retest scores obtained by patients with the same initial test score is given by the formula:

$$Z = \frac{Y - \hat{Y}}{\text{S. E.}_{\text{predic.}}} = \frac{83 - 120.10}{15.69} = -2.37$$

Consulting the table for the normal curve, this is equivalent to a percentile value of 1. Thus we see that only 1% of the schizophrenics in the "control" sample who started out with this IQ on the Wechsler, would have dropped as much on retest after four weeks as has our patient. This is consistent with our hypothesis that he has deteriorated.

Summary

This paper suggests simple statistical models for use in solving the following problems:

1. To estimate how large a discrepancy between any two test scores need be, for it to be "abnormally" large in the standardization population. That is, to estimate the frequency of occurrence in the standardization population of any given discrepancy between two test scores.
2. To establish how large a difference between two test scores need be, for it to be outside the range of differences produced solely by the errors of measurement of the two tests. That is to estimate how large a discrepancy must be for us to judge it a "measurable" difference.
3. To estimate how large a predicted change in score (following treatment, trauma, etc.) need be, to lie outside the range of changes found in a control group which has not been subjected to the intervening process.

References

1. Hamister, R. C. Test-retest reliability of the Wechsler Bellevue. *J. consult. Psychol.*, 1949, 13, 39-43.
2. Johnson, P. O. *Statistical Methods in Research.* New York: Prentice-Hall, 1949.
3. McNemar, Q. *Psychological Statistics.* New York: Wiley, 1949.
4. Slater, P. Interpreting Discrepancies. *Brit. J. Med. Psychol.* 1943, 19, 415-419.
5. U. S. Employment Service: *Guide to the use of General Aptitude Test Battery: Section III. Development.* Washington, D. C.: U. S. Dept. of Labor. 1955.
6. Wechsler, D. *The Measurement of Adult Intelligence.* (3rd ed.). Baltimore: Williams and Wilkins, 1944.