ASSESSMENT OF DIFFERENCES

Contents

If viewing on screen you can click on a contents item above to jump to the page the item is on

1. Formulas for assessing differences between scores on tests with large standardisation samples.

There are various methods for determining if a score is reliably or abnormally different from scores expected on the basis of a test's standardisation data. The purpose of this module is to present these methods and give examples of how to use them in practice.

The group with whom the individual is being compared by these formulas is of course the test's standardisation group.

1.1 The different sorts of differences

As these formulas were first suggested as a group of formulas for use in clinical assessment by Bob Payne and Gwynne Jones we will refer to them as the Payne and Jones formulas to distinguish them from others where this is appropriate. Here are the introductory remarks from the paper which first roused interest in the now commonly used formulas for assessing differences in score in the individual case.

From: Payne, R. W. and Jones, H. G. (1957) Statistics for the investigation of individual cases. Journal of clinical Psychology, 13, 115 - 121,

"Much of the work of clinical psychologists consists of making fairly routine 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 in order to assess test scores intelligently. There appear to be three main types of question which face clinical psychologists.

1. The abnormality of a difference between two scores.
This problem arises whenever the 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. 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 know that the scales are not perfectly correlated. Therefore many normal people would show discrepancies between the scales 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 use two tests which measure rather different traits. For example we might give a test of long term retention and a test of general intelligence. It may be the case that these tests have very low intercorrelation in the general population, so that quite large discrepancies between these scores are 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 that 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 '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 (pr 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. We need to know what proportion of individuals like our patient, of the same IQ on the first test who have not deteriorate 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 occurred. The practising psychologist will not have time to conduct the appropriate control experiment. Is there any other way of providing an approximate answer?"*

The paper then goes on to describe methods (see below) which assess discrepancies between test results by converting the difference seen between an individual's test scores into a Z score on the appropriate distribution of possible differences, i.e.

$$Z_{diff} = \frac{D - \overline{D}}{\sigma_{(D-\overline{D})}}$$

Where: Z_{diff} is a value to be looked up in tables for the normal curve

D is the difference observed between test scores

$\overline{D}$ is the mean difference between scores on the two tests

$\sigma_{(D-\overline{D})}$ is the standard deviation of the distribution of differences between two test scores

If we have converted our test scores to Z scores, all of the distributions of differences will have a mean of zero, but the standard deviations of the distributions will vary. These will in fact be as follows:

Problem	**Standard deviation of the appropriate distribution of differences when comparing discrepancies between Z scores on tests**
Abnormality of a difference	$\sqrt{2 - 2r_{xy}}$
Reliability of a difference	$\sqrt{2 - (r_{xx} + r_{yy})}$
Discrepancy between initial score and score on re-test	$\sqrt{1 - r_{xx}^2}$
Discrepancy between an obtained score on one test (Y) and the score on that test predicted by the obtained score on a different test (X)	$\sqrt{1 - r_{xy}^2}$

Reminder of the jargon

As would be expected from the excerpt from the Payne and Jones paper, the following terms are used as defined here.:

"Abnormal" means statistically rare. The criteria for rareness have usually followed the conventions of statistical hypothesis testing in psychology. So tests for the 'abnormality of a difference" estimate the frequency with which a difference of a given size would occur in the standardisation population, and, by inference, in the population at large.

Something which would be found in only 5% or 1% or 0.1% is taken to be 'abnormal. You would have to decide which of these cut-offs to use. Most people probably use the 5% level.

"Reliable" means 'unlikely to be due to measurement error.
You also have to remember that reliable differences are not necessarily "abnormal" differences.

So, let's look the formulas which can help us deal with the problems of :

Reliability of a difference between scores
Abnormality of a difference between scores
The significance of a change in score on re-test

We will also look at a fourth formula concerned with the abnormality of a difference between the obtained score on Test Y and the score on Test Y <u>predicted from the score on Test X.</u>

Reliability of a difference

This formula answers the question. “Is the difference I observe between two test scores a real difference or is it likely to be a mere chance difference?”

The formula, which helps answer, this question is:

$$Z_{diff} = \frac{Z_x - Z_y}{\sqrt{\left(2 - \left(r_{xx} + r_{yy}\right)\right)}}$$

Suppose that someone obtains a Z score of +1.5 on a visual memory test, and a Z score of + 0.5 on a test of auditory memory.

Suppose further that the reliability of the visual test is .70 and the reliability of the auditory test is .80.

Putting these values into the formula we get the following.

$$Z_{diff} = \frac{1.5 - .5}{\sqrt{2 - (.70 + .80)}} = \frac{1}{\sqrt{.50}} = \frac{1}{.71} = 1.41$$

This gives us a two-tail probability of 0.16 that the difference we have found is due to chance. By the usual standards of testing the significance of differences we would decide that there is no reason for rejecting the hypothesis that this difference is due to chance or errors of measurement.

Note that this formula does no more than tell you whether an observed difference is a real (not due to errors of measurement) one.

It does not tell you whether the difference is abnormal.

Abnormality of a difference.

This formula answers the question: "Is the difference I observe between two test scores sufficiently rare to set alarm bells ringing? Or is it fairly common to find a difference of this size?"

The formula which helps answer this question is:

$$Z_{diff} = \frac{Z_x - Z_y}{\sqrt{\left(2 - 2r_{xy}\right)}}$$

Let us continue with the example we used above. Suppose that someone obtains a Z score of +1.5 on a visual memory test, and a Z score of + 0.5 on a test of auditory memory.

Suppose further that the correlation between these two tests is 0.6.

Placing these values in our abnormality of a difference formula we get the following.

$$Z_{diff} = \frac{1.5 - .50}{\sqrt{\left(2 - \left(2 \times .60\right)\right)}} = \frac{1}{\sqrt{.80}} = \frac{1}{.89} = 1.12$$

A difference of this size (one standard deviation or larger) would be found in about 26 percent of the population.

We would therefore conclude that it is not abnormally large.

Difference between a score on first testing and a re-test score..

This formula is used in a situation where we have reason to believe that somebody might have changed, e.g., deteriorated intellectually; or become less depressed; or become more anxious, etc.

The initial score is used to predict what the re-test score would be expected to be.

Two additional "error" factors would need to be taken into account in assessing the significance of any observed change. These are

1. Regression effect

Because the test will not have perfect reliability, it follows from the formula for predicting one score from another, that the score on the second testing would be expected to be nearer to the mean than the score on the first occasion.

You will recall that the formula is:

$$\hat{Z}_y = r_{xy} Z_x$$

Unless $r_{xy} = 1.0$, the predicted Z_y will always closer to the mean than Z_x.

The formula itself takes care of the regression effect for you.

2. practice effect

On some tests, mostly tests of cognitive abilities practice effects are known to occur. The score on re-test tends to be higher as a result of doing the first test.

The formula as it stands does not take account of practice effects. (A modified formula which does will be given later.)

$$Z_{diff} = \frac{Z_{x_2} - r_{x_1 x_2} Z_{x_1}}{\sqrt{1 - r^2_{x_1 x_2}}}$$

Where:

r_{x1x2} is the test-retest reliability coefficient of the test

Continuing with our example, let's suppose that we retested our client on the visual memory test. Suppose further that the score on retest was a Z of 0.5. We have stated above that the reliability of this test was 0.7.

Putting these values into our formula gives us the following.

$$Z_{diff} = \frac{0.5-(0.7\times 1.5)}{\sqrt{1-(.7\times .7)}} = \frac{-0.55}{.71} = -0.77$$

Consulting tables for the normal curve we find that a difference as large as this would be expected to occur in about 44 percent of cases, and that a difference as large as this <u>in this direction</u> in 22 percent of cases.

If practice effects are known for the test in question the formulas should be modified to deal with them. This is done bit subtracting the practice effect from the second test score. If we are dealing in Z scores the correction to be subtracted from the second score will be

$$\frac{practice\ effect}{\sigma_x}$$

The modified formula thus becomes

$$Z_{diff} = \frac{\left(Z_{x_2} - \frac{pe}{\sigma_x}\right) - r_{x_1x_2} Z_{x_1}}{\sqrt{1-r^2_{x_1x_2}}}$$

A fourth formula

The prediction formula can also be used to assess the abnormality of a difference between scores on two different tests.

Its use in this context differs from that of the abnormality of a difference formula given above. The formula above is concerned wit how common a difference of a give size is in the standardisation/general population.

If we use the prediction formula to assess the abnormality of a difference, it will tell us how common a given difference between scores on two different tests is amongst people who all obtained the same score on the first test.

Let's try this variation on our running example. We supposed that someone obtained a Z score of +1.5 on a visual memory test, and a Z score of + 0.5 on a test of auditory memory, and that the correlation between these tests was 0.6.

The formula for the abnormality of a difference between test scores if we use this method is:

$$Z_{diff} = \frac{Z_y - r_{xy} Z_x}{\sqrt{1 - r_{xy}^2}}$$

Putting the values into this formula we get

$$Z_{diff} = \frac{0.5 - (1.5 \times .6)}{\sqrt{1 - (.6 \times .6)}} = \frac{-.4}{.8} = -.5$$

So, about 62 percent (two-tail) of people who obtained a Z score of 1.5 on the first test would a show a bigger difference between scores than this. And about 31 percent (one-tail) of people with a Z score of 1.5 on the first test would show a bigger drop in score on the second test.

1.2 Reliability of the difference between scores on two different tests

When you have two scores on two different tests and you know the reliability coefficients of each of both of the tests, then you can calculate the probability of obtaining a difference of this size between the two scores, using the formula below.

$$Z_{difference} = \frac{Z_x - Z_y}{\sqrt{2-(r_{xx}+r_{yy})}} \qquad \textbf{(1)}$$

Where Zx = z-score of the score on first test
Zy = z-score of the score on second test
r_{xx} = reliability coefficient of the first test
r_{yy} = reliability coefficient of the second test

Example:

On a test of non-verbal intelligence a individual scores at the 75th percentile, while on a test of verbal intelligence their score is at the 25th percentile. The reliability coefficients of the two tests are .80 and .84. Is this difference possibly due to errors of measurement on the two tests?

Z of 75th percentile = 0.67,
Z of 25th percentile = - 0.67

Using formula (1) above

$$\frac{0.67 - -0.67}{\sqrt{2-(0.80+0.84)}} = \frac{1.33}{0.60} = 2.22$$

The probability of obtaining a difference this large or larger as result of measurement error on the two tests is double the proportion of the normal curve that lies beyond a z of 2.22 (the difference could have been negative or positive so we need to use a two tailed test of significance to include both tails of the normal curve). That probability is 0.0264 (0.0132 x 2), therefore there is only about a 3% chance that the difference is due only to measurement error. As such there is good reason for supposing that the difference is significant.

1.3 The reliability of a difference between scores on the same test

When you have two scores from two different individuals on the same test and you know the reliability coefficient of that test, then you can calculate the probability of obtaining a difference of this size between the two scores, using the formula below.

$$Z_{difference} = \frac{Z_{x1} - Z_{x2}}{\sqrt{(2 - 2r_{xx})}} \qquad \textbf{(2)}$$

Where $Zx\ 1$ = z-score of the score for the first person
$Zx2$ = z-score of the score for the second person
r_{xx} = reliability coefficient of the test

Example

Two individuals are being considered as candidates for entry into a training program, a prerequisite for which is that they have good mathematical ability. The two individuals are given a test of mathematical ability, obtaining scores of 121 and 123. The test has a mean of 100, a Standard deviation of 15 and a reliability of 0.95. They both exceed the course entry cuttoff value of 120, but there is only one place left in the course and so only one can be selected. Should you recommend the candidate with the higher score? In order to decide you need to know if the differences between the two scores are a real reflection of differences in ability or are they likely to have arisen due to chance variation due to test measurement error.

Z of a score of 123 = (123 – 100)/15 = 1.53,
Z of a score of 121 = (121 – 100)/15 = 1.40

Using formula (2) above

$$z = \frac{1.40 - 1.53}{\sqrt{2 - 1.9}} = \frac{-0.13}{0.32} = -0.41$$

The probability of obtaining a difference this large or larger as result of measurement error on the tests is the proportion of the normal curve that lies beyond a z of 0.41. That probability is 0.6818 (two-tails). Therefore there is about a 68% chance that the difference is due only to measurement error, therefore it is unlikely to be a real difference and you would conclude they have the same level of ability as measured by the test.

1.4 The abnormality of a difference between two scores

If you know the correlation between two tests and you want to compare scores from the two tests on the same individual, then you can apply the formula below.

$$Z_{difference} = \frac{Z_x - Z_y}{\sqrt{2 - 2r_{xy}}} \qquad \textbf{(3)}$$

Where Zx = z-score of the score on first test

Zy = z-score of the score on second test

r_{xy} = correlation coefficient between the two tests

Example

A child is assessed with the Stanford Binet -IV (SB5) test, obtaining a score of 114 and the Wechsler Individual Achievement Test – II (WIAT-II) obtaining a score of 101. The correlation between these two tests is reported as 0.80 in the SB5 Technical Manual (p 93). How common (or rare) would such a difference in score between these two tests be?

Z of a score of 114 = (114 – 100)/15 = 0.93, Z of a score of 101 = (101 – 100)/15 = 0.07

Applying formula (3) above:

$$Z_{diff} = \frac{0.93 - 0.07}{\sqrt{2 - 1.60}} = \frac{0.86}{0.63} = 1.37$$

The probability of obtaining a difference this large or larger on these tests is the proportion of the normal curve that lies beyond a z of 1.37. That probability is 0.1707 (two-tails). Therefore about 17% of the population would exhibit a difference such as this.

1.5 The abnormality of a difference between a predicted and an obtained score

Instead of comparing the difference between two scores, you may want to compare a score with a predicted score (predicted from a score on another test). Doing this takes into account regression to the mean effects, and this makes the comparison more accurate. The following formula does this.

$$Z_{difference} = \frac{Z_y - r_{xy} Z_x}{\sqrt{1 - r_{xy}^2}} \qquad \textbf{(4)}$$

Where Zx = z-score of the score on first test

Zy = z-score of the score on second test

r_{xy} = correlation coefficient between the two tests

Example

A child was recently assessed as having a WISC-III Full Scale IQ score of 92. As a second opinion you assess the child using the Stanford Binet–IV (SB5) and obtain a Full Scale IQ score of 80. The SB5 technical manual reports the correlation between the WISC-III and the SB5 to be .84 (p 87). How common or rare would this difference be?

Z of a score of 92 = (92 – 100)/15 = -0.53,

Z of a score of 80 = (80-100)/15 = -1.33

Applying Formula (4) above:

$$Z_{diff} = \frac{-0.53 - (0.84 * -1.33)}{\sqrt{1 - 0.84 * 0.84}} = \frac{-1.65}{0.54} = -3.06$$

The probability of obtaining a difference this large or larger on these tests is the proportion of the normal curve that lies beyond a z of 3.06. That probability is 0.0022 (two-tail). Therefore about 1in 2000 of the population would exhibit a difference in score like this on these two tests. It is reasonable to conclude that the difference is significant.

1.6 The abnormality of a change in score on the same test between two test occasions

Clinically, we often want to compare two scores obtained on the same test by the same individual on two different occasions. Such as when we want to know if a person has clinically improved or deteriorated or has otherwise changed in some clinically meaningful way, that can be measured with a test. In such situations we should compare the second score with a score predicted by the first score, so that we take into account regression to the mean effects. Also with many tests of ability, there are practice effects. That is if a person has been tested before, it often the case this primes them to obtain a better score second time round, particularly if the time interval between test occasions is short. Where appropriate these practice effects also need to be taken into account. The following formula should be used to calculate the abnormality (as a probability) of a change in score on the same test.

$$Z_{diff} = \frac{\left(Z_{x_2} - \frac{pe}{\sigma_x}\right) - r_{x_1x_2} Z_{x_1}}{\sqrt{1 - r_{x_1x_2}^2}}$$

Or

$$Z_{diff} = \frac{Z_{(x_2 - pe)} - r_{x_1x_2} Z_{x_1}}{\sqrt{1 - r_{x_1x_2}^2}} \qquad \textbf{(5)}$$

Where $Z_{(x_2 - pe)} =$ z-score of (the score on the second occasion – practice effect)

$Z_{x_1} =$ z-score of the score on the first occasion

$r_{x_1x_2} =$ test-retest reliability coefficient of the test

$pe =$ Practice Effect

$\sigma_x =$ Standard Deviation of the test

Note: the second formula is computationally simpler to use

Example1

A child is tested on an intelligence test at age 5 years obtaining a score of 140. At age 6 years they are again tested with the same test and obtain a score of 160. The mean of the test is 100, the standard deviation is 15, the test-retest reliability over 1 year for this age group is .76 and the practice effect over one year is 7 points.

X2 – PE = 160 – 7 = 153

Z of a score of 153 = (153 - 100)/15 = 3.53

Z of a score of 140 = (140 – 100)/15 = 2.67

Applying formula (5) above:

$$Z_{diff} = \frac{3.53 - 0.76 * 2.67}{\sqrt{1 - 0.58}} = \frac{1.50}{0.65} = 2.31$$

The probability of obtaining a difference this large or larger on this test is the proportion of the normal curve that lies beyond a z of 2.31. That probability is 0.0209 (two-tail). Therefore about 2% of the population would exhibit a difference this large or larger on this test and about 1% would exhibit an increase this large or larger. As such it is reasonable to conclude the child's IQ has probably improved.

Example 2:

You are a psychologist providing CBT for depression to a client. As part of your pre-intervention assessment, the client completes a Depression Inventory. They obtain a score of 25. After eight sessions of CBT you want to assess how much the client has improved and you ask the client again completes the same Depression Inventory, this time they obtain a score of 14. You want to know, how significant is this change of score on the Depression Inventory?

First you need to gather some data about the test to use the appropriate formula. You consult the test manual. The test-retest correlation reported in the manual is 0.93. The manual also reports means and standard deviations for several groups. There are norms for several groups reported in the manual, "Depressed Inpatients, College Students and Depressed Outpatients" You decide the "Outpatients" group (N=500) reported in the manual most closely matches your client in clinical description and so choose to use data from this group. For the outpatients group the mean is 22.45 and the standard deviation is 12.75. The manual summarises studies of practice effects of the Depression Inventory and concludes that they are negligible. Generally only cognitive or ability tests have practice effects. Therefore there is no practice effect to take into account in your calculations.

Z of a score of 14 = (14 - 22.45)/12.75 = - 0.66,

Z of a score of 25 = (25 – 22.45)/12.75 = 0.02

Applying the formula (5)

$$Z_{diff} = \frac{-0.66 - 0.93 * 0.02}{\sqrt{1 - 0.93 * 0.93}} = \frac{-0.68}{0.36} = -1.89$$

The probability of obtaining a drop this large or larger on these tests is the proportion of the normal curve that lies beyond a z of -1.89. That probability is 0.0294 (one tail). Therefore only about 3% of the population would exhibit a drop in score like this on the Depression Inventory. As such it is reasonable to conclude the client's Depression Inventory score has significantly reduced.

***Note* that because we were testing for a reduction in score a one-tail test was used. The issue of when to use a one-tail or two-tail probability will be discussed in more detail later**

1.7 Which formula should be used when?

Which formula should I use? The answer depends upon the clinical question that you are trying to answer and the availability of information such as reliability coefficients and correlations between two tests. The reliability of a test (r_{xx}) is generally available in the test manual. Correlations between tests may be in some manuals, especially if a test is used with another test often, or they can be found in published papers where they exist. The Database of Facts on www.psychassessment.com.au contains some of this information.

The table below can be used as a guide as to which formula can be used to answer different kinds of clinical questions.

	Two test scores (x and y) **Different tests or scales**	**Two Test occasions (x1 and x2)** **Same test or scale**
Reliability Asks: Is the difference due to test measurement error?	$\frac{Z_x - Z_y}{\sqrt{2-(r_{xx}+r_{yy})}}$ compares the same individual on two different test	$\frac{Z_{x_1} - Z_{x_2}}{\sqrt{(2-2r_{x_1x_2})}}$ compares two different individuals on the same test or compares the same individual on the same test on two different occasions
Abnormality Asks: How probable (or rare) is this difference in the population?	$\frac{Z_x - Z_y}{\sqrt{2-2r_{xy}}}$ compares the same individual on two different test	
Abnormality With predicted score Asks: How probable (or rare) is this difference in the population?	$\frac{Z_y - r_{xy}Z_x}{\sqrt{1-r_{xy}^2}}$ compares the same individual on difference between their obtained score on a test and the score predicted on the same test from another test	$\frac{Z_{(x_2-PE)} - r_{x_1x_2}Z_{x_1}}{\sqrt{1-r_{x_1x_2}^2}}$ compares the same individual on the same test on two different occasions and takes into account the practice effect (PE), if no practice effect PE= 0

These formulas work quite well with most reasonably standardised tests. But sometimes standardisation samples are small and more exact techniques may be needed. These are covered in the advanced topic on Differences on www.psychassessment.com.au .

2. One tail or two tail probabilities?

In clinical practice we are sometimes concerned with a null hypothesis i.e. there is nothing really wrong with this person, there has been no change in this person, etc. In fact, with the more 'serious' diagnoses we could be in human rights trouble if we did not start with an open mind.

But in clinical problems we are also at other times trying to decide whether the patient has become worse, or become better. This particular question is a one-tail test one. But the more general question has there been any change in this person is a two tail one

The most appropriate action is to report exactly what you have found and are trying to convey.

X% of patients with depression would have worsened this much or more.
Y% of those with dementia would have been expected to improve this much or more. And so on.

In general, testing any difference which becomes apparent, after administering a test, should be two-tailed.

A specific query about change in a particular direction can and should be one tailed. But the criterion chosen can also depend on the relative weight of the consequences of the decision in one direction as opposed to another.

For a one-tail test, descriptive statements such as X percent would have improved more than this or Y percent would have deteriorated less than this is probably the best option

Generally a two-tail test is the most appropriate, if in doubt use a two-tail test, but if the clinical question is similar to “has a depressed patients mood improved?” or “has a patient's dementia progressed?” etc the case for a one tail test is clear.

You can also report both: e.g.:” X percent would show a difference this large and X/2 percent would show a difference this large in this direction.”

3. What level of probability is significantly abnormal?

The z-score formula presented in this paper, are used to determine the abnormality of a difference as a probability. That is what proportion of the population if chosen at random would exhibit a difference of this size or larger. The decision that a given difference is abnormal is a clinical decision and depends largely on the clinical assessment situation. In some situations a 10% probability of a difference is abnormal in others it may need to be below 5% or below 1%.

In the last resort this decision can be made by using the 5% significance convention. In the statistical analysis of psychological data the conventional levels of significance have been set at a probability of < 0.05 or lower. By somewhat indirect and possibly stretched analogy some have called a score abnormal if it is at the 95^{th} percentile or higher or at the 5^{th} percentile or lower. The 5% criteria is not a magical number. In the end, statistical abnormality is matter of degree, not of kind.

4. When is a single score abnormal?

In the last resort this decision is made by convention. In the statistical analysis of psychological data the conventional levels of significance have been set at a probability of $<$ 0.05 or lower. By somewhat indirect and possibly stretched analogy some have called a score abnormal if it is at the 95^{th} percentile or higher or at the 5^{th} percentile or lower.

With tests which have been standardised on large samples and which have normally distributed scores we can easily use tables for the normal curve to see if a given score meets our criterion for abnormality. For example the WAIS-III has a mean of 100 and standard deviation of 15. If an assessment with the WAIS-III finds that someone has an IQ of 64, then this translates to a z-score of -2.4, which in turn translates to a probability of 0.0082 or 0.8% that a member of the population chosen at random will have an IQ score on the WAIS-III of 64 or lower. This would be considered abnormal.

But there are many tests in use which have relatively small normative groups, and some in which the nature of the population distribution of scores is unknown. See the section on Differences in the Advanced Topics section of www.psychassessment.com.au , that describes possible methods for assessing the abnormality of scores or of differences between scores in such circumstances.