ASSESSMENT OF DIFFERENCES 2

Contents

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

1. Tchebycheff's Inequality

For our purposes, this inequality can be described as being concerned with the probability that a score will differ from the mean by more than a certain amount.

It will apply to any continuous distribution of test scores.

Tchebycheff proved the following:

$$probability\left(\left|X-\mu\right| \geq b\right) \leq \frac{\sigma^2}{b^2}$$

This can be roughly translated (for our purposes) as:

The probability that a score will deviate from the mean by '*b*' units is always less than or equal to the variance of the test divided by b^2

Again, let Z-scores make life easier for us. In Z scores the formula becomes:

$$probability\ (Z_{score} \geq k) \leq \frac{1}{k^2}$$

This translates into:

The probability of getting a Z score of equal to or greater than a stated value (k), is less than or equal to 1 divided by k^2.

So, let's suppose we have a test with a small amount of normative data and for which we cannot be certain of the shape of the population distribution.

The mean of the test is 15 and the standard deviation is 5. Somebody scores 30 on this test. Can we be certain that this score (3 standard deviations above the mean) is abnormal if our criterion of abnormality is scoring higher than 95 percent of people?

Entering the values into the equation gives us the following:

$$probablility\ (Z_{score} \geq 3) \leq \frac{1}{9} = 0.11$$

Thus p could be as high as 0.11, so we cannot say that the score meets our criterion.

However, if we know for sure that the distribution of scores is uni-modal and symmetrical, the formula becomes:

$$probability(Z_{score} \geq k) \leq \frac{4}{9}\left(\frac{1}{k^2}\right)$$

Thus the probability of getting a Z-score of 3 would be equal to or less than .44 x .11 which is equal to or less than 0. 0484

So, given the extra information about the distribution, we could conclude that the score met our criterion for abnormality.

2. Grubbs Test for an Outlier

Based on Grubbs, F. E. (1969) Procedures for detecting outlying observations in samples. *Technometrics*, **11**, 1 –21

This test is useful when we want to test the hypothesis that an individual's score is too discrepant for it to be likely that that individual belongs to that particular group.

Grubb's test is used when the suspect individual obtains either the highest or the lowest score in the group

Suppose that there is a rare clinical condition for which we have a test with only a fairly small amount of data for the rare diagnostic group. We have data for, say, only 11 people suffering from the disorder. There is no reason to believe that the data are normally distributed.

We suspect that a new patient might be suffering from that condition and compare their score with that of the normative group.

The new patient obtains a core higher than the highest score in the standardisation group, and high scores indicate lack of that particular pathology.

Is the patient's score significantly different from that of the normative group?

Grubb's Test can help us answer the question.

The formula is:

$$T = \frac{\left(X - M_{x+}\right)}{\sigma_{x+}}$$

The + subscript in the case of both the mean and the standard deviation is there to indicate that both should be based on the original group plus the suspect score. In addition the standard deviation should be an estimate of the population standard deviation (i.e. divided by N – 1, not just N)

Suppose the scores were 2, 3, 4, 4, 5, 5, 5, 6, 6, 9, and 10. The new patient obtains a score of 14.

If we now add this score to the previous set we can use Grubb's Test to test the hypothesis that all of the scores could represent a random sample drawn from the same underlying distribution.

If we do the appropriate sums we find the following values

$$T = \frac{(14-6)}{3.30} = 2.42$$

Looking this value up in the table below, we find that with n = 12, the new patient's score is significantly higher (p<. 05) than the 'standardisation' group's scores.

The test is also applicable to other sorts of 'score'

Suppose we had a group of 5 people who showed the following differences in their scores on a visual – an auditory digit recall test: 4, 5, 6, 8, and 9. A newly tested individual has a discrepancy score of 1. Is this different from the other scores? These scores with new score added to them have a mean of 5.5 and an estimated population standard deviation of 2.88. The calculation for Grubb's T is therefore

$$T = \frac{4.5}{2.88} = 1.56$$

Consulting the table we find that the one-tail critical value for significance at the .05 level is 1.82, and for two-tail significance at the .05 level is 1.89. We conclude therefore that the new score does not differ significantly from the others.

An excellent calculator for Grubb's Test can be found at:

http://graphpad.com/quickcalcs/grubbs2.cfm

The paper by Grubbs, referred to above, also gives details of a number of other tests for detecting outliers.

Grubb's Test: One-tail critical values for T
(Adapted from Grubbs, F. E. (1969) Procedures for detecting outlying observations in samples. *Technometrics*, **11**, 1 –21.)

Number of scores n	5 percent significance level	2.5 percent significance level	1 percent significance level
3	1.15	1.15	1.15
4	1.46	1.48	1.49
5	1.67	1.71	1.75
6	1.82	1.89	1.94
7	1.94	2.02	2.10
8	2.03	2.13	2.22
9	2.11	2.21	2.32
10	2.18	2.29	2.41
11	2.23	2.36	2.48
12	2.29	2,41	2.55
13	2.33	2.46	2.61
14	2.37	2.51	2.66
15	2.41	2.55	2.71
16	2.44	2.59	2.75
17	2.47	2.62	2.79
15	2.50	2.65	2.82
19	2.53	2.68	2.85
20	2.56	2.71	2.88
21	2.58	2.73	2.91
22	2.60	2.76	2.94
23	2.62	2.78	2.96
24	2.64	2.80	2.99
25	2.66	2.82	3.01
30	2.75	2.91	
35	2.82	2.98	
40	2.87	3.04	
45	2.92	3.09	
50	2.96	3.13	
60	3.03	3.20	
70	3.09	3.26	
80	3.14	3.31	
90	3.18	3.35	
100	3.21	3.38	

3. The Crawford, Howell and Garthwaite modification of the standard Payne and Jones Formula for the abnormality of a difference between two scores.

(Crawford, J. R., Howell, D. C., and Garthwaite, P. H. (1998) Payne and Jones revisited: Estimating the abnormality of test score differences using a modified paired sample t-test. *Journal of clinical and experimental Neuropsychology*, **20**, 898-905)

These authors argue persuasively that, when control/normative samples are small the ***t*** distribution is more suitable for assessing significance than the use of the standard normal distribution.

Accordingly they have proposed the following variant of the Payne and Jones formula for the abnormality of a difference.

$$t = \frac{Z_x - Z_y}{\sqrt{(2 - 2r_{xy})\left(\frac{N_2 + 1}{N_2}\right)}}$$

where:

N_2 is the number of cases in the group with whom the individual is being compared.

So, if we wished to compare a patient's discrepancy between 2 scores with the discrepancies obtained by a control group of 15 people, and the patient's Zx was 1.8 and Zy was 1, and the correlation between X and y was .8 we would need to calculate the value:

$$t = \frac{1.8 - 1.0}{\sqrt{(2 - 1.6)\left(\frac{16}{15}\right)}} = 1.22$$

Looking this value up in the t tables we find a one-tailed probability of 0.12 for the probability of a difference between X and Y of this size.

Had we used the Payne and Jones Formula, the estimated probability of a difference as large as this would have been 0.10.

4. The Crawford and Howell modification of the formula for assessing the abnormality of a difference between a predicted and an obtained score

(Crawford, J. R. and Howell, D. C. (1998) Regression equations in clinical neuropsychology: An evaluation of statistical methods for comparing predicted and obtained scores. *Journal of clinical and experimental Neuropsychology*, **20**, 755-762)

Crawford and Howell have pointed out that the use of the traditional formula for the standard error of estimate used in clinical prediction problems is perhaps not accurate enough for use with regression equations based on small normative samples.

So instead of the traditional formula:

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

with $\sqrt{1-r_{xy}^2}$ or, in raw scores, $\sigma_y\sqrt{1-r_{xy}^2}$ as the standard error of estimate and the $Z_{difference}$ looked up in Tables for the standard normal distribution, they recommend a more exact alternative for the standard error of estimate with the result assessed by reference to the *t* distribution.

For raw scores the standard error of estimate becomes:

$$\sigma_y\sqrt{1-r_{xy}^2} \times \left(\sqrt{1+\frac{1}{N}+\frac{\left(X_0-\overline{X}\right)^2}{\sigma_x^2\left(N-1\right)}}\right)$$

Where:

X_0 signifies the new score that we are using to predict Y from X.

Needless to say there is a Z-score equivalent formula, which is:

$$\sqrt{1-r_{xy}^2} \times \left(\sqrt{1+\frac{\left(1+Z_o^2\right)}{N}} \right)$$

As will be clear from inspecting this formula, as N increases, the value of the correction term shrinks. For example, for a Z of 2.0 and N = 20, the value of the correction would be:

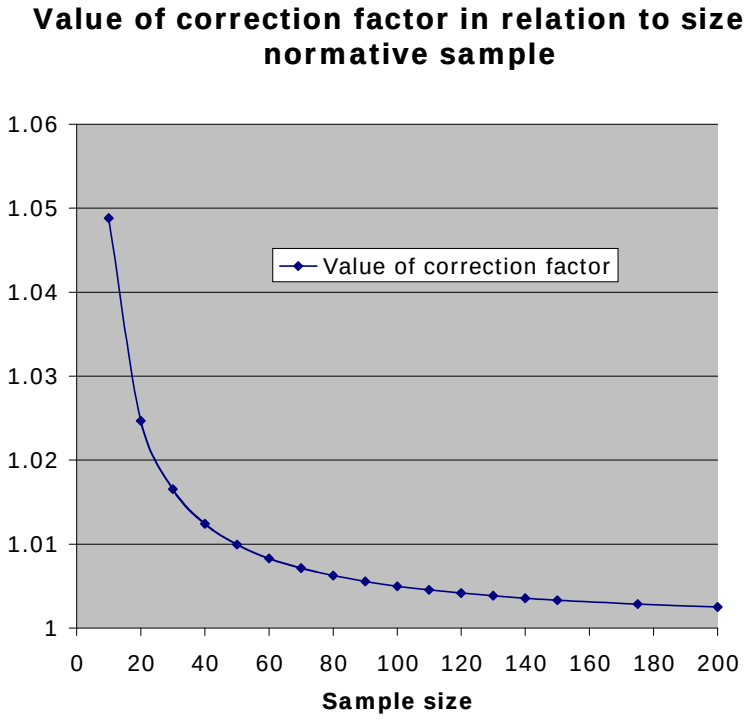
$$\sqrt{1+\frac{1+2^2}{20}} = 1.12$$

While if N was 50 the value of the correction would only be:

$$\sqrt{1+\frac{1+2^2}{50}} = 1.05$$

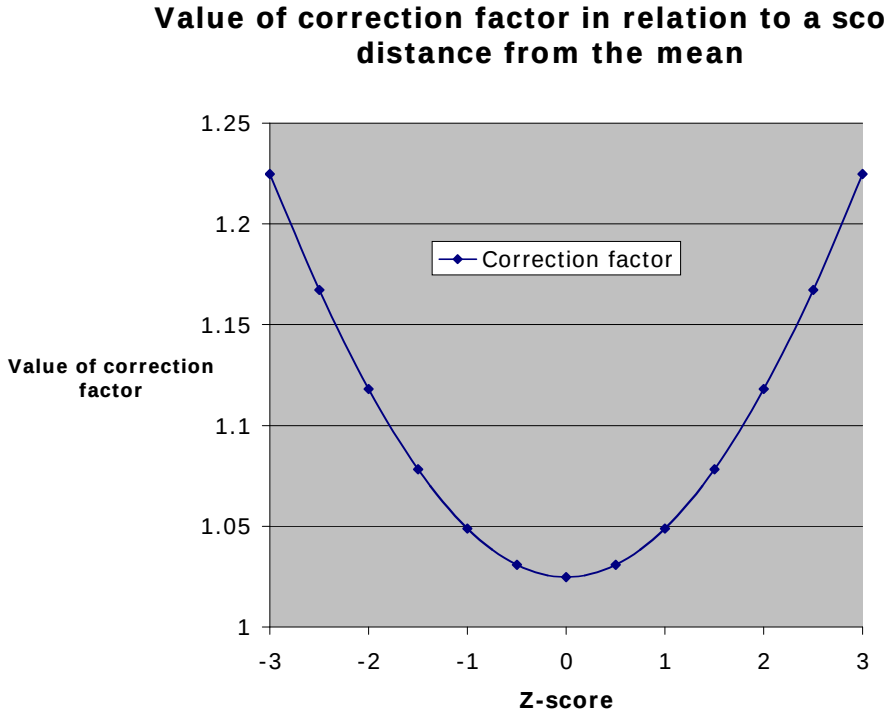
And for someone obtaining the mean score the value of Z would be zero and the correction factor would simply be 1 + 1/N

The graph below shows the size of the correction factor for a mean score for different sizes of N.



And, as was implicit above, the further from the mean a Z-score is, the greater will be the correction factor. This is shown below for a sample size of 20.

So the formula refinement corrects both for sample size and for the effect of a score's distance from the mean.



To use the test for a difference between an obtained and a predicted score, the full modified Z-score formula is:

$$t = \frac{Z_y - r_{xy} Z_x}{\sqrt{1 - r_{xy}^2} \times \left(\sqrt{1 + \frac{\left(1 + Z_x^2\right)}{N}} \right)}$$

The degrees of freedom for the *t* – test will be N – 2.

Information, papers and calculators based on Crawford's work can be found at:

http://www.abdn.ac.uk/~psy086/dept/psychom.htm

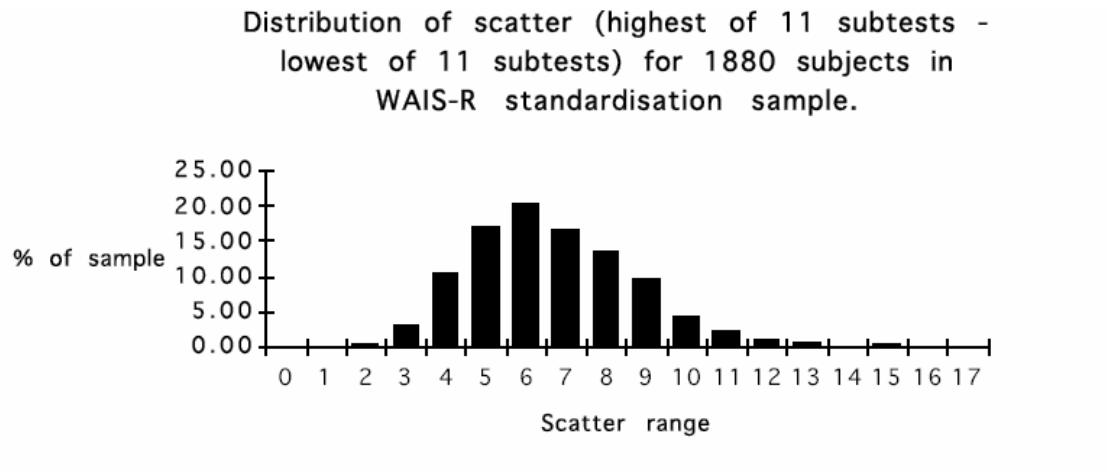
5. Empirical assessment of differences

One obvious way to assess the abnormality of differences is simply to tabulate their frequency from the test's standardisation data.

Most major test manuals now provide appropriate data.

Here for example is a chart showing the distribution of scatter on the WAIS-R. Scatter is, of course, the difference between the highest and lowest scores an individual obtains on a test, or a number of tests.

On the WAIS-R, mean scatter for the Verbal Scale is 4.67; on the Performance Scale is 4.71; and for the Full Scale is 6.66. The distributions of scatter ranges are skewed. Across the full scale of 11 subtests, approximately 30% of people show a scatter range of 7 points or more; 20% of 8 or more; 10% of 9 or more; and 5% of 10 or more.

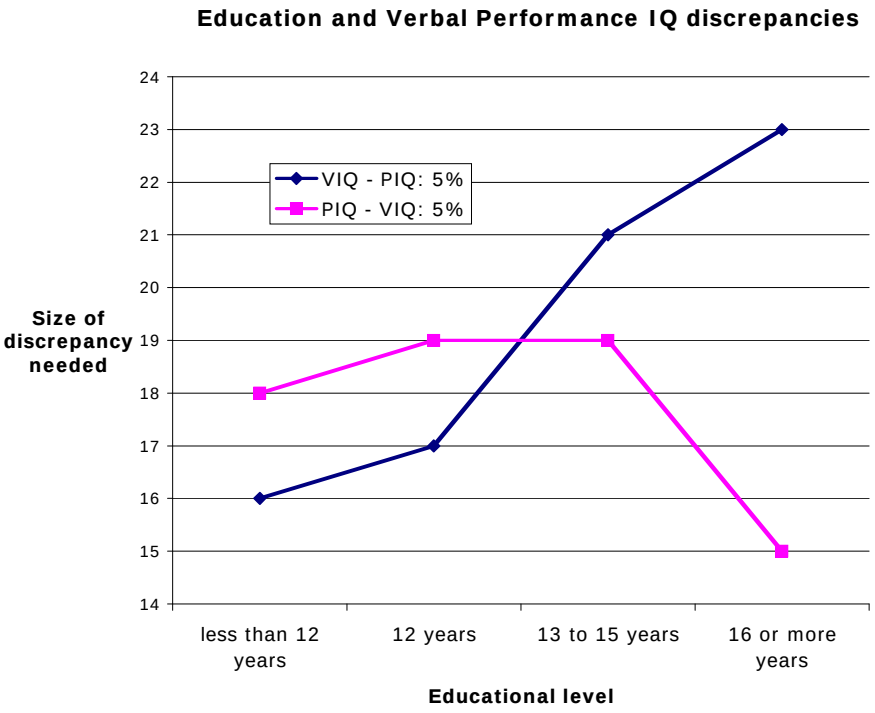


In addition to the materials in the manual, some test publishers have made the standardisation data available to researchers who have produced a number of more detailed analyses.

For example, Dori and Chelune (2004) reported on the frequency with which discrepancies of given sizes occur in groups with different amounts of education, within and between WAIS and WMS scales.

For instance, the graph below shows how large Verbal IQ minus Performance IQ, and Performance minus Verbal IQ discrepancies have to be to be found in five percent or less of people with the stated amount of education.

Thus for people with 16 or more years of education the 95th percentile for Verbal minus Performance discrepancy is not reached until the difference is 23 points, while for those with less than 12 years of education the 95th percentile for this difference is reached at 16 points, and so on.



he graph is based on part of Table 1 in Dori, G. A. and Chelune, G. J. (2004) Education-stratified base-rate information on discrepancy scores within and between the Wechsler Adult Intelligence Scale – Third Edition and the Wechsler Memory Scale – Third Edition. *Psychological Assessment*, **16**, 146-154

We do not know why the values in the graph depart from what we would expect from use of the formula. Perhaps, it is because of small sample sizes, perhaps it is because higher verbal ability is a determinant of staying longer in the educational system, perhaps it is for other reasons.

But these data do suggest that larger Verbal minus Performance IQ differences might be more frequent the more education a person has undergone, and such discrepancies should thus be treated with some caution in highly educated groups.

The general moral of this tale is that norms should be as specific as is reasonably possible to the person being assessed.