Quantitative Aspects of Psychological Assessment

An Introduction

Philip Ley

Chapter One

The summation sign and the rules of summation

1. The summation Sign

It is frequently necessary in statistical and psychometric calculations to take the sum of a number of values. The symbol used to indicate this operation of adding up a group of numbers is a capital Greek Sigma - $\sum$.

However, the instruction 'to take the sum of' is rather vague without an indication of what it is that is to be summed. It is necessary to have a system of notation to specify precisely which values are to be summed. Let us suppose that we have a set of four scores:

2, 4, 6, 8,

and that we let *X* be a general symbol for any one of these scores. The set of scores now consists of four *X's* which are 2, 4, 6 and 8. If we now assign a subscript to each of the *X's*, we can assign an *X* which a given subscript to each score thus:

$$X_1 = 2; \quad X_2 = 4; \quad X_3 = 6; \quad X_4 = 8.$$

In the case of four scores the subscripts will naturally run from 1 to 4, but there are usually more than four scores involved, and so it is desirable to have a generalised notation, so that we can apply the system to any group of scores. A general symbol for the number of scores is *N*. Any collection of scores will consist of *N* scores. If we want a general reference to a single score without specifying exactly which we can use the subscript *i*. Thus X is the *i*th score. Consider the following set of scores:

$$X_1 = 6; \quad X_2 = 7 \quad X_3 = 8; \quad X_4 = 9; \quad X_5 = 10.$$

What is the value of X_i when *i* = *N*? The answer is 10. There are five scores, therefore *N* = 5. X_i when *i* = *N* must be X_5, which is the symbol for the fifth score, which is 10. In similar fashion the value of X_1 where *i* = 2 is 7; where *i* = 1 it is 6 and so on.

These symbols are often used in connection with the summation sign to indicate exactly which scores are to be summed.

For example:

$$\sum_{i=1}^{N} X_1 = \text{take the sum of all scores from } X_1 to X_N$$

i.e. $X_1 + X_2 + X_3 + \ldots X_N$

The symbols above and below the summation sign are called the limits of the summation. The value of *i* under the summation sign tells you where to start the addition, and the values above the summation sign tells you where to stop. The starting and stopping places can be anywhere in the set of scores.

$$\sum_{i=2}^{4} X_i = \text{take the sum of scores 2 to 4}$$

i.e. $X_2 + X_3 + X_4$

Often when there is no danger of confusion the symbol $\sum X$
Is used without subscripts or limits. This should be read as though it was:

$$\sum_{i=1}^{N} X_i = \sum X = \text{Take the sum of all the numbers}$$

i.e. $X_1 + X_2 + X_3 + \ldots X_N$

$\sum X$ will be used frequently in this book, but always check to see if there are any subscripts with it.

Sometimes more than one summation sign is used. Suppose that our scores are classified into groups, and let us use the symbol *J* for the number of groups, and n_j for the number of cases in the *j*th group. A particular X_i will now be found in the *j*th group, so we can put a double subscript under the *X*, to make it clear that we are talking about the *i*th score in the *j*th group, hence X_{ij}. Suppose now that we want to find the total score for one of the groups. We will need to sum all of the X_{ij} in that group and there will be n_jX_{ij}'s. So the instruction to sum all the scores in a group can be written:

$$\sum_{i=1}^{n_j} X_{ij} = \text{take the sum of all the scores in a group.}$$

If we now want to sum these totals of groups to find the grand total for all scores we can write:

$$\sum_{j=1}^{J}\sum_{i=1}^{n_j} X_{ij} = \qquad \text{find the groups' totals}$$

and add them all together.

It will not be necessary in the following chapters to use more than double summation, but as many summation signs as necessary may be used.

The use of brackets is also important. If confronted with instructions inside a bracket always follow these instructions before following the instructions outside the brackets.

$$\sum_{i=1}^{N} X_1^2 + X_2^2 + X_3^2 + \ldots X_N^2$$

but

$$\left(\sum_{i=1}^{N} X_i\right)^2 = \left(X_1 + X_2 + X_3 + \ldots X_N\right)^2$$

which is not the same as $\sum_{i=1}^{N} X_{i/}^2$

Similarly

$\sum_{j=1}^{J}\sum_{i=1}^{n_j} X_{ij}^2 =$ Square every number in a group and find the total of the squared numbers, repeat for all groups and then sum the group totals.

While

$\sum_{j=1}^{J}\left(\sum_{i=1}^{n_j} X_{ij}\right)^2 =$ Find the total score for a group, square this total, repeat for each group, then sum the squared group totals.

So

$$\sum_{j=1}^{J}\sum_{i=1}^{n_j} X_{ij}^2 \neq \sum_{j=1}^{J}\left(\sum_{i=1}^{n_j} X_{ij}\right)^2$$

Problems

A. Given the following set of scores

$X_1 = 4$; $X_2 = 5$; $X_3 = 6$; $X_4 = 7$; $X_5 = 8$; $X_6 = 9$; $X_7 = 10$;
$X_8 = 11$; $X_9 = 12$; $X_{10} = 13$; $X_{11} = 14$; $X_{12} = 15$.

(1) What is the value of N?
(2) What is the value of X_i when I = 6?
(3) What is the value of X_i when I = N?

(4) What are the values of:

(a) $\sum_{i=1}^{4} X_i$ (b) $\sum_{i=10}^{N} X_i$ (c) $\sum_{i=5}^{8} X_i$

B. In the following groups indicate which values will be the same as another.

(1) (a) $\left(\sum_{i=1}^{N} X^2\right)$ (b) $\left(\sum_{i=1}^{N} X\right)^2$ (c) $\sum_{i=1}^{N} X^2$

(2) (a) $\left(\sum_{j=1}^{J}\sum_{i=1}^{n_j} X_{ij}\right)^2$ (b) $\sum_{j=1}^{J}\left(\sum_{i=1}^{n_j} X_{ij}\right)^2$

(c) $\sum_{j=1}^{J}\sum_{i=1}^{n_j} X_{ij}^2$ (d) $\sum_{j=1}^{J}\left(\sum_{i=1}^{n_j} X_{ij}^2\right)$

Answers

A. (1) 12; (2) 9; (3) 15; (4) (a) 22; (b) 42; (c) 38.
B. (1) (a) and (c); (2) (c) and (d).

2. Some Rules of Summation

On several occasions later in this book, proofs will be presented which require knowledge of some of the rules of summation. In this section the rules will be stated and proofs of the rules provided. The proofs are easy and well within the competence of any reader of this book. The reader is therefore urged to read the proof as well as the rule.

Summation Rule 1: *The sum of the sums of two or more Variables is equal to the sum of their Summations.*

i.e. $$\sum_{i=1}^{N}\left(X_i+Y_i+Z_i\right)=\sum_{i=1}^{N}X_i+\sum_{i=1}^{N}Y_i+\sum_{i=1}^{N}Z_i$$

Proof

(1) $$\sum_{i=1}^{N}\left(X_i+Y_i+Z_i\right)=\left(X_1+Y_1+Z_1\right)$$

$$+\left(X_2+Y_2+Z_2\right)+\left(X_3+Y_3+Z_3\right)$$

$$+\ldots\left(X_N+Y_N+Z_N\right)$$

(2) Removing the brackets leaves

$X_1 + Y_1 + Z_1 + X_2 + Y_2 + Z_2 + X_3 + Y_3 + Z_3 \ldots + X_N + Y_N + Z_N$

(3) This equals all the *X*'s plus all of the *Y*'s plus all of the *Z*'s.

(4) Which is the same as

$$\sum_{i=1}^{N}X_i+\sum_{i=1}^{N}Y_1+\sum_{i=1}^{N}Z_i$$

Summation Rule 2: *The sum of a constant times the values of a variable is equal to the constant times the sum of the variable.*

i.e. $\sum_{i=1}^{N}\left(cX_1\right)=c\sum_{i=1}^{N}X_i$ where *c* is a constant.

Proof

(1) $\sum_{i=1}^{N}cX_1 = cX_1 + cX_2 + cX_3 ... + cX_N$

(2) As everything is multiplied by *c* this can be written as: c(X_1 + X_2 + X_3…X_N)

(3) The term inside the brackets is $\sum_{i=1}^{N}X_i$ therefore

C(X_1 + X_2 + X_3 …+ X_N) + c$\sum_{i=1}^{N}X_i$

(4) The brackets can be removed to give c$\sum_{i=1}^{N}X_i$

Summation Rule 3: *The sum of a constant taken N times is The constant times N.*

i.e. $\sum_{i=1}^{N}c = Nc$

Proof

(1) $\sum_{i=1}^{N}c = c + c + c ... c$

(2) It can be seen that *N* constants are added together. This is the same as taking the constant *N* times which equals *Nc.*

Summation Rule 4: *The sum of the values of a variable plus A constant, is equal to the sum of the Values of the variable plus N times the constant.*

i.e. $$\sum_{i=1}^{N}(X_i + c) = \sum_{i=1}^{N} X_i + Nc$$

Proof

(1) $\sum_{i=1}^{N}(X_i + c) = \sum_{i=1}^{N} X_i + \sum_{i=1}^{N} c$ (by Summation Rule 1).

(2) $\sum_{i=1}^{N} c = Nc$ (by Summation Rule 3).

(3) So we obtain $\sum_{i=1}^{N} X_i + Nc$

Summation Rule 5: *The sum of the values of a variable minus a Constant is equal to the sum of the values Of the variable minus N times the constant.*

i.e. $$\sum_{i=1}^{N}(X_i - c) = \sum_{i=1}^{N} X_i - Nc$$

Proof

This can be left as an exercise.

Problems

A. Simplify the following expression:

$$\sum_{i=1}^{N}\left(X_i+Y_i-c-d\right)$$ where *c* and *d* are constants.

B. Check your answer using the following data:

Subject	*X Score*	*Y Score*
(1)	1	5
(2)	2	6
(3)	3	7
(4)	4	8
	c = 2	d = 3

C. In the above example what is the value of

(a) $\sum_{i=1}^{N} d$ (b) $\sum_{i=1}^{N} cX_i$ (c) $\sum_{i=1}^{N}\left(X_i+Y_i\right)$?

D. Are the answers you gave to question C consistent with the rules of summation given above?

Answers

A. $\sum_{i=1}^{N}X_i+\sum_{i=1}^{N}Y_i-N(c+d)$

B. $\left(X_i+Y_i-c-d\right)$ for subject (1) is 1; for (2) is 3; for (3)is 5 and for (4) is 7. The sum of these is 16.

$$\sum_{i=1}^{N}X_i=10;\sum_{i=1}^{N}Y_i=26;N=4;c+d=5$$

Inserting these values in A gives 16 which is the same as for:

$$\sum_{i=1}^{N}\left(X_i+Y_i-c-d\right)$$

C. (a) 12. (b) 20. (c) 36.

The mean, the variance, and the standard deviation

1. The Mean and other Measures of Central Tendency

The arithmetic mean is the average of a set of numbers. It can be symbolised as *M* and its formula is:

$$\text{Arithmetic Mean} = M = \frac{\sum_{i=1}^{N} X}{N} = \frac{\sum X}{N} \qquad (2:1)$$

It is the most important of three measures of central tendency. The other two are the median and the mode. The mode is defined simply as the value which occurs most frequently. The median is the value below which exactly fifty percent of cases fall, and above which are exactly fifty percent of cases. It is the point which splits the set of scores into two equal parts.

Problems

Find the mean, mode and median of the following sets of scores.

A. 1, 2, 2, 3, 4, 5, 6.

B. 10, 11, 12, 14, 15, 15.

C. 7, 7, 8, 8, 10, 10.

Answers

A. Mean 3.3; mode 2; median 3.

B. Mean 12.8; mode 15; median 13
(by convention half way between the two mid-most scores when there is an even number of scores).

C. Mean 8.3; mode 8; median 8.

From this point onwards subscripts will be used only when necessary.

Returning now to the formula for the mean it can be shown that:

$$\sum X = NM \qquad (2:2)$$

Proof

(1) $NM = N\left(\dfrac{\sum X}{N}\right).$

(2) $= \dfrac{N\sum X}{N}.$

(3) The *N*'s cancel one another out leaving $\sum X$.

This general principle that the sum of a set of values equals *N* times the mean of that set will be useful at several later points.

It is also true that:

$$\sum \left(X - M\right) = 0 \qquad (2:3)$$

Proof

(1) $\sum(X-M)=\sum X-\sum M$ by Summation Rule 1.

(2) Further by Summation Rule 3, as the mean is a constant:

$\sum M = NM$, so (1) becomes $\sum X - NM$.

(3) But we have just shown in equation (2:2) that $NM = \sum X$ so we obtain $\sum X - \sum X = 0$.

The value $X_i - M_x$, the deviation of a score from the mean, is called a deviation score and is sometimes symbolised as x_i. Similarly $Y_i - M_y$ is symbolised as y_i. As demonstrated in (2:3)

$$\sum x = 0; \qquad \sum y = 0$$

2. The Variance

The variance is a measure of dispersion. It tells us something about the scatter of scores around the mean. It is defined as the mean squared deviation from the mean, and symbolised by a small sigma squared - σ^2. Its formula is:

$$\text{Variance} = \sigma_x^2 = \frac{\sum(X-M)^2}{N} \qquad (2:4)$$

or using *x* for X – M

$$\sigma_x^2 = \frac{\sum x^2}{N} \qquad (2:5)$$

It follows from this formula that:

$$\sum x^2 = \sum (X - M)^2 = N\sigma_x^2 \qquad (2:6)$$

(2:6) is obtained from (2:5) by multiplying both sides of the equation by *N*.

Another variant of the formula for the variance is:

$$\sigma_x^2 = \frac{\sum x^2}{N} - M^2 \qquad (2:7)$$

Proof

(1) $\sigma_x^2 = \sum (X - M)^2 / N.$

(2) $= \sum (X^2 + M^2 - 2XM) / N.$

(3) Using Summation Rules 1 and 3 this becomes:

$$\left(\sum X^2 + NM^2 - 2\sum XM\right) / N$$

(4) but we have shown in (2:2) that $\sum X = NM$ so we can write:

$$\sigma_x^2 = \left(\sum X^2 + NM^2 - 2NMM\right) / N$$

(5) but $2NMM = 2NM^2$ so (4) becomes:

$$\sigma_x^2 = \left(\sum X^2 - NM^2\right) / N$$

(6) Dividing by *N* this gives:

$$\sigma_x^2 = \frac{\sum X^2}{N} - M^2$$

The numerator (top part) of equation (2:4) for the variance is the sum of squared deviations from the mean. This sum is usually called the sum of squares.

$$\text{Sum of squares} = SS = \sum (X - M)^2 \text{(2:8)}$$

An alternative formula for this value is:

$$SS = \sum X^2 - \frac{(\sum X^2)}{N} \qquad \text{(2:9)}$$

Proof

(1) In the proof of (2:7) at (5) it has been shown that

$$\sum (X - M)^2 = \sum X^2 - NM^2$$

(2) But $NM^2 = N\left(\frac{\sum X}{N}\right)\left(\frac{\sum X}{N}\right)$

(3) Multiplying this becomes: $\frac{N\sum X \sum X}{N^2}$

(4) Dividing numerator and denominator by *N* gives:

$$\frac{(\sum X^2)}{N}$$

(5) Substituting this in (1) we obtain:

$$\sum (X - M)^2 = \sum X^2 - \frac{(\sum X^2)}{N}$$

1. *The Standard Deviation*

The standard deviation is the square root of the variance and is symbolised by a small Greek sigma - σ. Its formula is the square root of any of the formulae for the variance, e.g.

$$\sigma_x = \sqrt{\frac{\sum x^2}{N}} \qquad (2:10)$$

The mean, the variance and the standard deviation are important in psychometrics because of their relationships to the normal curve. These relationships will be discussed in the next chapter.

Problems

Given the following set of scores:

1, 2, 3, 4, 5, 6, 7.

Find:

A. The mean.
B. The variance.
C. The standard deviation.
D. $\sum(X-M)$.

Answers

A. 4; B. 4; C. 2; D. 0.

The standard normal distribution

1. *Distributions in General*

If we plot frequency distributions of data, the distributions obtained can vary in a number of ways. Three of these ways are:

(1) modality;
(2) skew;
(3) kurtosis.

Modality refers to the number of peaks in a distribution. Three distributions varying in modality are shown in Figure 3.1, unimodal, bimodal and trimodal. The trimodal distribution has been labelled 'polymodal', a general term for distributions with more than one mode.

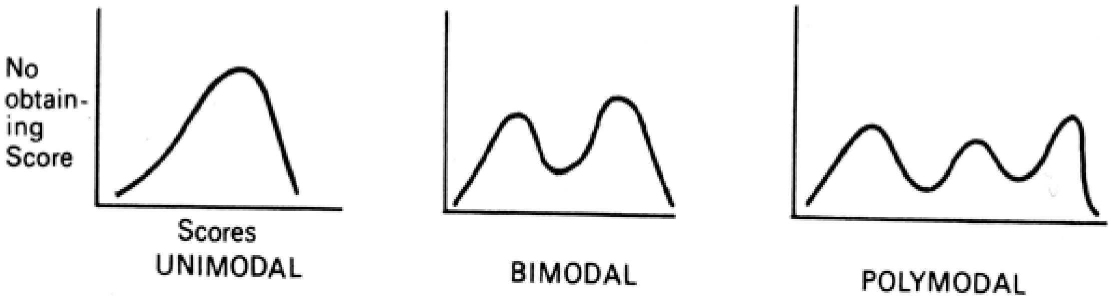


Figure 3.1 Distributions with differing numbers of modes

Unimodal distributions can differ from one another in terms of skew. A skewed distribution is asymmetrical and has a tapering tail at one end. The tapering part is the skewed part and if it is in the direction of low scores the distribution is negatively skewed, while if it is in the direction of high scores the distribution is positively skewed. Skewed distributions are shown in Figure 3.2.

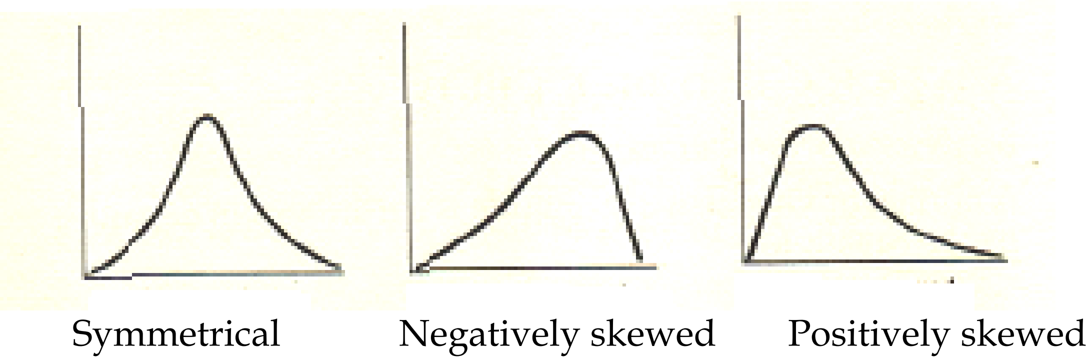


Figure 3.2 Symmetrical and skewed distributions

Distributions also differ in kurtosis or degree of peakedness, tall narrow distributions being lepto-kurtic and short broad ones platy-kurtic. Figure 3.3 shows distributions differing in kurtosis.

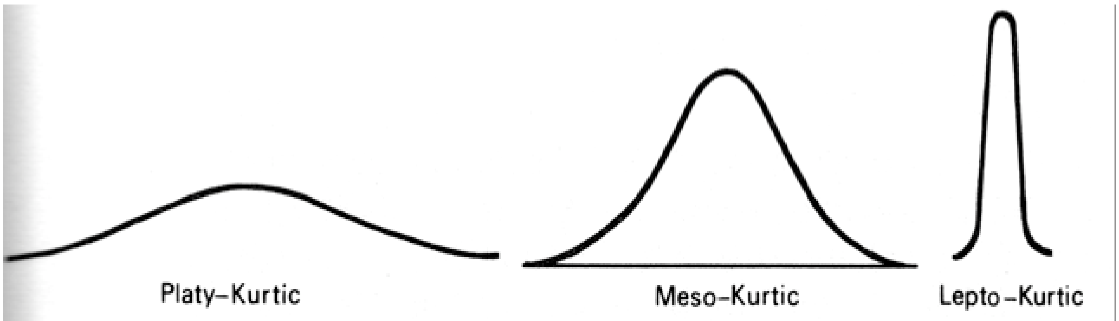


Figure 3.3 Different degrees of Kurtosis in frequency distributions

Problems

Given the following distributions of percentages of subjects obtaining the indicated scores:

Scores	*I*	*II*	*III*	*IV*	*V*
70-79	30	10	5	5	
60-69	40	30	10	10	
50-59	20	10	15	15	5
40-49	5	10	20	20	20
30-39	5	30	25	20	50
20-29		5	15	15	20
10-19		5	10	10	5
0- 9				5	

A. Which of the symmetrical distributions is the more platy-kurtic?

Which of all the distributions is, or are:

B. negatively skewed;
C. bimodal;
D. positively skewed;
E. unimodal.

Answers

A. IV;
B. I;
C. II;
D. III;
E. I, III, IV, V.

2. *Normal Distributions*

A normal or Gaussian distribution is a distribution described by the equation:

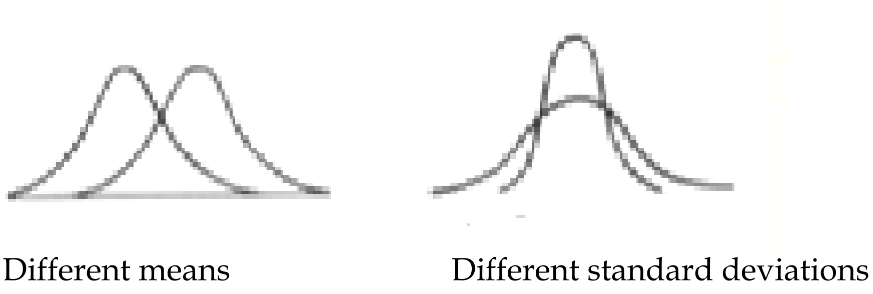
$$Y = \frac{1}{\sigma\sqrt{2\pi}} e^{-(X-M)^2/2\sigma^2} \qquad (3:1)$$

where:

Y = Proportion of cases at a given point.
π= 3.1416
e = 2.718
σ = Standard deviation
M = Mean.

The equation may be a little off-putting at first glance but close inspection reveals that the mean and standard deviation are important parts of the formula. In fact the formula for any normal distribution is the same except-t for these two values.

Figure 3.4 shows some normal distributions which differ in these characteristics.



Different means

Different standard deviations

Figure 3.4 Normal distributions differing in standard deviation or mean

All normal distributions have the same general shape but they can differ tremendously in degree of kurtosis, yet amongst the things that all normal distributions have in common is the fact that the mean, the median and the mode fall in the same place. In a normal distribution the mean = the median = the mode. Further in all normal distributions the range M ± 3σ includes nearly all cases.

Given the mean and standard deviation of a normal distribution the probability of occurrence can be worked out for any value. It is therefore possible to prepare tables, which give these probabilities, but these would differ from one distribution to another because of differences in the numerical value of the means and standard deviations.

To circumvent this problem it is necessary to find a common unit of measurement into which any score could be converted so that one table will do for all normal distributions. This common unit is found in the standard score or Z score.

3. *Standard Scores or Z scores*

Z scores are scores converted into the number of standard deviations that the scores are from the mean of their distribution.

$$Z = \frac{X - M}{\sigma} = \frac{x}{\sigma} \qquad (3:2)$$

It can be seen from (3:2) that, to find a *Z* score, the difference between a score and the mean is divided by the standard deviation of the scores.

A *Z* score of +2.0 therefore means that the original score was 2 standard deviations above the mean.

A *Z* of –3.5 means that the original score was three and a half standard deviations below the mean.

Problem

Convert the following set of scores into *Z* scores:

1, 2, 3, 4, 5, 6, 7.

Answer

Step 1. $$M = \frac{\sum X}{N} = \frac{28}{7} = 4$$

Step 2. $$\sigma = \sqrt{\frac{\left(-3^2\right) + \left(-2^2\right) + \left(-1^2\right) + 1^2 + 2^2 + 3^2}{7}} = 2$$

Step 3. $$Z = \frac{X - M}{\sigma} = \frac{1-4}{2}; \frac{2-4}{2}\ etc.$$

$$= -1.5;\ -1.0; -0.5; 0; +0.5; +1.0; +1.5$$

The mean of a set of *Z* scores is zero. As we will need to use means of several statistics in later sections we will symbolise means by putting a bar over the symbol representing the statistic.

Thus

$\overline{Z_x} =$ mean *Z* score for the X's,

$\overline{X}$ = M_x, $\overline{\sigma}$ = mean standard deviation and so on.

$$\overline{Z} = 0 \qquad (3{:}3)$$

Proof

(1) $\overline{Z} = \frac{\sum Z}{N}.$

(2)

$$\frac{\sum Z}{N} = \frac{\sum(X-M)}{\sigma} / N$$

(3) But it has been shown that

$\sum(X-M) = 0$ (See 2:3 for details)

So $\frac{\sum Z}{N} = \frac{0}{\sigma} / N = 0$

The variance of a set of *Z* scores is 1.0, and as the square root of 1.0 is itself 1.0, the standard deviation of a set of *Z* cores is also 1.0.

$$\sigma_z^2 = \sigma_z = 1.0 \qquad (3{:}4)$$

Proof

(1) $\sigma_z^2 = \frac{\sum\left(Z - \overline{Z}\right)^2}{N}$ (This is just the usual formula for the variance with Z's instead of X's.)

(2) As $\overline{Z} = 0$, (from 3:3)

$$\sigma_z^2 = \frac{\sum Z^2}{N}$$

(3) But $Z = \frac{X - M}{\sigma} so \sum Z^2 = \frac{\sum\left(X - M\right)^2}{\sigma^2}$

(4) However if $\frac{\sum\left(X - M\right)^2}{N} = \sigma^2 \, then \sum\left(X - M\right)^2 = N\sigma^2$

(5) Using the information from (3) and (4) we obtain:

$$\sum Z^2 = \frac{\sum\left(X - M\right)^2}{\sigma^2} = \frac{N\sigma^2}{\sigma^2} = N$$

(6) So $\sigma_z^2 = \frac{\sum X^2}{N} and \sum Z^2 = N.$ Therefore

$$\sigma_z^2 = \frac{\sum Z^2}{N} = \frac{N}{N} = 1.0$$

We now have a distribution with a mean of zero and a standard deviation of 1.0, and we can easily convert any score into a Z score by use of formula (3:2).

4. *The Standard Normal Distribution*

The standard normal distribution is the normal distribution with a mean zero and a standard deviation of one, and a total area under its curve of 1.0. The meaning of the area under the curve will become clear in the examples. It is simply the proportion of cases. Tables are available for the normal distribution, and can be found in almost any elementary statistics text. Unfortunately the data selected for inclusion may differ from text to text. Tables in the text books give one or more of the following values:

(a) The proportion of cases falling in the area between the mean and a given *Z* score, and/or

(b) The proportion of cases falling beyond a given *Z* score. This is called the proportion in the smaller area, and/or

(c) The proportion in the larger area cut-off by a given *Z* score.

Detailed instructions and examples are given below for the use of each of these types of tables. Tables 3:1, 3:2 and 3:3 give selected values from each of the different types. In all of these tables the left hand column is a *Z* score. Before each table there is a diagram indicating the areas involved.

An example of the first sort of table is given below in Table 3:1. This table gives the area between the mean and a given *Z* score.

Figure 3.5 shows the area of the curve lying between the mean and a *Z* of –1.0.

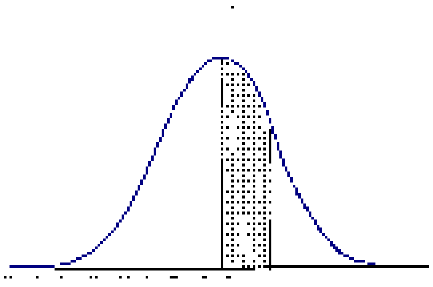


Figure 3.5 Proportion of cases lying between the mean and z score (shaded area)

TABLE 3.1 PROPORTION OF CASES LYING BETWEEN THE MEAN AND A GIVEN STANDARD SCORE

x/σ = Z	*Proportion of area of Curve between M and Z*
0.00	.0000
0.10	.0398
0.20	.0793
0.50	.1915
1.00	.3413
2.00	.4772
3.00	.49865

A number of problems can be solved using this table.

(i) To find the proportion of cases scoring above a given point.

(a) If *Z* is positive, the proportion scoring above a given point is given by .5000 minus the proportion lying between the mean and the value of *Z*.

Example If Z is +1.00, the proportion of cases lying between *Z* and the mean is .3413, therefore the proportion above this point is .5000 minus .3413 = .1587.

(b) If *Z* is negative, the proportion scoring above a point is given by .5000 plus the proportion lying between the mean and the *Z* value.

Example If Z is –1.00 the proportion of cases lying between this value and the mean, will be the same as that lying between the mean and a Z scores of + 1.00, because the normal curve is perfectly symmetrical. Thus the required proportion will be .3413 + .5000 - .8413.

(ii) To find the proportion of cases falling below a certain point.

(a) If Z is positive the proportions of cases falling below a given point will be equal to .5000 plus the proportion of cases between the mean and that *Z* score.

Example The proportion of cases falling below a *Z* score of -2.00 is equal to .5000 + .4772 = .9772.

(b) If *Z* is negative the proportion of cases falling below a point will equal .5000 minus the area between the mean and that *Z* score.

Example The proportion of cases falling below a Z score Of – 2.00 equals .5000 - .4772 = .0228.

(iii) To find the proportion of cases falling between two given points.

(a) If both *Z* scores have the same sign, i.e. if both are positive or both are negative, the proportion falling between the two points will be the proportion lying between the mean and the higher *Z* minus the proportion lying between the mean and the lower *Z*.

Example What proportion of cases lie between a *Z* score of +1.00 and a *Z* score of + 2.00? The proportion between the mean and +2.00 = .4772, while the proportion between the mean and + 1.00 = .3413, thus the proportion lying between the two is .4772 - .3413 = .1359.

(b) If the *Z* scores have unlike sign, i.e. one is positive and the other is negative, then the proportion lying between them is the sum of the proportions lying between the mean and each *Z* score.

Example What proportion of cases lie in the range between +1.00 and -.50? The proportion between the mean and + 1.00 = .3413, while that between the mean and - .50 equals .1915. Therefore the proportion of cases falling in the range between +1.00 and -.50 = .3413 + .1915 = .5328.

The second type of table described above gives the proportion of cases lying further away from the mean than a given *Z* score. Some values from such a table are given in Table 3:2. Again a visual aid is provided in Figure 3.6 which shows the proportion of cases falling in the area beyond a *Z* of +1.0.

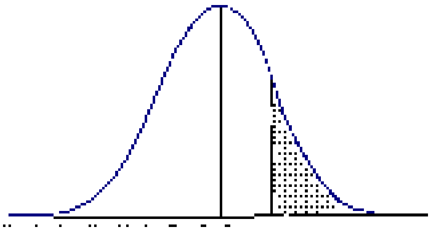


Figure 3.6 Proportion of cases in the smaller portion

TABLE 3:2 PROPORTIONS IN THE SMALLER PORTION OF THE CURVE FOR DIFFERENT VALUES OF *Z*

x/σ = Z	*Proportion falling in area further away from the mean than the specified Z score*
0.00	.5000
0.10	.4602
0.20	.4207
0.50	.3085
1.00	.1587
2.00	.0228
3.00	.00135

Just as in Table 3:1 the Z score values start at .00, which is the mean, but values in the body of this table can be seen to be .5000 minus the corresponding value in Table 3:1.

The rules for using this Table 3:2 are therefore different.

(i) To find the proportion of cases scoring above a given point.

(a) If *Z* is positive the value in the table opposite that *Z* will be the proportion scoring above that point.

Example If Z is + 1.00 what proportion of cases will score above that value? Opposite a Z of 1.00 is the proportion .1587, which is the proportion of cases scoring above that value.

(b) If *Z* is negative the proportion of cases scoring above the value will be 1.0000 minus the proportion opposite *Z* in the table.

Example If Z is –1.00 what proportion of cases will obtain a higher score? The proportion of cases opposite 1.00 in the table is .1587 therefore the proportion scoring above *Z* score of 1.00 will be 1.0000 - .1587 = .8413.

(ii) To find the proportion of cases scoring below a given point.

(a) If *Z* is positive the proportion will be 1.00 minus the proportion opposite the value of *Z* in the table.

Example The proportion of cases falling below a *Z* of +2.00 is equal to 1.00 - .0228 = .9772.

(b) If Z is negative the proportion will be equal to the value in the table.

Example The proportion of cases falling below a Z score of –2.00 equals .0228.

(iii) To find the proportion of cases falling between two specified points.

(a) If both *Z*′s have the same sign, the proportion lying between them will be the difference between the proportions in the table corresponding to the *Z*′s.

Example What proportion of cases lies in the range between a Z score of + 1.00 and a Z score of +2.00? Reference to Table 3:2 shows that a proportion of .1587 obtains higher scores than a *Z* of 1 and .9228 obtains higher scores than a *Z* of 2. By the rule the proportion lying in the range

$$Z_1 - Z_2 = .1587 - .0228 = .1359.$$

(b) If the *Z*′s have unlike sign, the proportion lying between them will equal .5000, Z_1 plus .5000 minus the proportion corresponding to Z_2.

Example What proportion of cases lies in the range between a Z of + 1.00 and a *Z* of –0.50? The proportion corresponding to a *Z* of +1.0 is .1587 and the proportion corresponding to a *Z* of –0.50 is .3085.

> Subtracting each of these from .5000 gives .3413 and .1915. Adding these gives .5328.

The third sort of table gives the proportion of cases lying in the larger area. An example of this type of table is given in Table 3:3. Figure 3.7 shows the proportion of cases in the larger area when *Z* is +1.0.

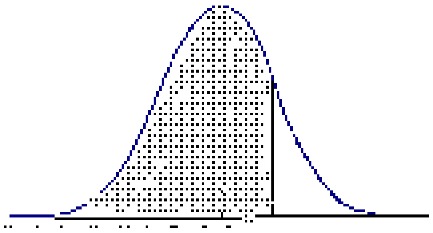


Figure 3.7 Proportion of cases in larger area when *Z* = +1.0

TABLE 3.3 AREAS IN THE LARGER PORTION OF THE NORMAL CURVE FOR DIFFERENT VALUES OF *Z*

$x/\sigma = Z$	*Area in the larger portion*
0.00	.5000
0.10	.5393
0.20	.5793
0.50	.6915
1.00	.8413
2.00	.9772
3.00	.99865

Problems

1. Write the rule for finding the number of cases falling above a given point.

2. Write the rule for finding the number of cases falling below a given point.

3. Write the rule for finding the number of cases falling between two points.

Answers

1. To find the proportion of cases falling above a given point:

   (a) If *Z* is positive subtract the value in the table corresponding to *Z* from 1.

   Example If *Z* is +1.00 what proportion of cases will score above that value? The proportion corresponding to a *Z* of +1.00 is .8413, this proportion subtracted from 1.000 leaves .1587.

   (b) If *Z* is negative the proportion opposite *Z* in the table gives the proportion of cases above that point.

   Example If *Z* is –1.00 what proportion of cases will score above that point? The answer directly from the table is .8413.

2. To find the proportion of cases falling below a point:

   (a) If *Z* is positive this can be read directly from the able. *Example* The proportion of cases falling below a *Z* of +2.00 is .9772.

   (b) If *Z* is negative the proportion falling below that point will be 1.0000 minus the proportion in the table corresponding to *Z*.

   Example What proportion of cases fall below a *Z* of -2.00? The answer will be 1.0000 - .9772 which equals .0228.

3. To find the proportion of cases falling between two given points:

   (a) If the *Z*'s have the same sign the answer is obtained by finding the difference between the proportions corresponding to the two *Z*'s.

   Example The proportion of cases lying in the range between a *Z* of +1.00 and a *Z* of +2.00 will equal .9772 - .8413 = .1359.

   (b) If the *Z*'s have unlike sign, the proportion will be the proportion corresponding to Z_1 minus .5000 plus the proportion corresponding to Z_2 minus .5000.

   Example The proportion of cases falling between a *Z* of +1.00 and a *Z* of -.50 will be .8413 - .5000 plus .6915 - .5000 = .5328.

It is hoped that during the solving of these problems the reader has gained insight into the methods for using the three types of table. If not the rules provided can be followed mechanically until insight dawns. There will be plenty of further opportunities to use the tables, especially in the next chapter.

Test Scales and norms

1. *Types of Test Score in Common Use*

The most frequent methods of reporting test results appear to be:

Percentiles
T Scores
I.Q's
Sten Scores

This chapter will describe the nature of each of these types of scale, and give the method of converting scores of one type into scores of another. At the present time it is possible to have data for an individual on a number of tests each of which gives its results differently. It is therefore necessary to be able to compare one type of scale with another.

2. *Percentiles*

Percentiles may be taken as points on a measuring scale below which a stated percentage of cases fall. Thus below the 75th percentile will fall 75 percent of cases, below the 10th percentile 10 percent, and so on. The 50th percentile is the median, and in the case of a normal distribution it is the mean and mode as well. In constructing percentile norms the first step is to tabulate the frequency distribution of the scores. When this has been done a cumulative frequency distribution is worked out. This gives the cumulative total of cases at each score level working usually from the lowest score upwards. Once this has been obtained the percentiles corresponding to the scores are worked out by the formula:

$$\text{Percentile} = \left(\frac{cfB = .5f}{N} \right) \times 100 \qquad (4:1)$$

where:

*cf*B = cumulutive frequency of the score below the one for which the percentile is being calculated;

f = frequency of the score whose percentile is being calculated;

N = total number of cases.

The complete procedure is illustrated below in Table 4:1.

TABLE 4:1 OBTAINING PERCENTILE NORMS FROM FREQUENCY DISTRIBUTIONS

Score	*Frequency of Score(f)*	*Cumulative Frequency of Score (cf)*	*cfB + .5f*	*Percentile (cfB + .5f)/N* × 100
50	2	20	19	95
49	3	18	16.5	82.5
48	4	15	13	65
47	5	11	8.5	42.5
46	3	6	4.5	22.5
45	2	3	2	10
44	1	1	.5	2.5

It should be emphasized, if it is not immediately apparent, that percentile norms would never be worked out on such a small number of cases. The process can also be worked in the opposite direction. If we want to find the score corresponding to a given percentile we use the formula:

Score corresponding to a percentile

$$= XLL + W\left(\frac{Np - cfB}{f}\right) \qquad (4:2)$$

where:

XLL = the lower limit of the class interval, in this case the scores are in units, so *XLL* = X -.5.
e.g. *XLL* for *X*50 = 49.5, for *X*37 = 36.5 etc.

W = Width of class interval in units.
e.g suppose scores were classed 90-94, 95-99, 100-104. *W* would be 5. In Table 4:1 *W* = 1.

Np = Number of cases times the proportion indicated by the percentile. Referring this value to the cumulative frequency column enables us to find the interval in which the value corresponding to the percentile will lie.

cf B and *f* are as above in formula 4:1.

As an example suppose with the data in Table 4:1 we wanted to find the 75th percentile, we would take the following steps:

Step 1. The value of *Np* will be 20 x .75 = 15. So we need the interval containing the 15th case.

Step 2. This corresponds to a score of 48, and in this interval are 4 cases, so *f* = 4.

Step 3. *cfB* = 11.

Step 4. The lower limit of the interval containing the 75th percentile is 47.5.

Step 5. Putting these together we get:

$$47.5 + 1\left(\frac{15-11}{4}\right) = 48.5 \qquad = \left\lfloor XLL + W\left(\frac{Np - cfB}{f}\right)\right\rfloor$$

3. *T Scores*

T scores are normally distributed scores with a mean of 50 and a standard deviation of 10. If the distribution of obtained scores is normal then *T* scores can be worked out directly by:

(a) converting each score to a *Z* score;

(b) multiplying the *Z* by 10; then

(c) adding or subtracting (depending on the sign of the *Z*) to or from 50. This will be the *T* score.

$$T = 10\left(\frac{X - M}{\sigma_x}\right) + 50 \qquad (4:3)$$

If the scores are not normally distributed the *T* scores will have to be calculated from percentiles. This procedure incidentally has the effect of normalizing the distribution of scores. Table 4:2 contains the raw scores and percentiles from Table 4:1. Two columns have been added. One of these gives the *Z* values corresponding to percentiles. These were obtained from tables for the normal curve, and the other gives *T* score values.

TABLE 4:2 CALUCLATION OF *T* SCORES FROM RAW SCORES

Score	*Percentile*	*Z Score*	*T Score (50 + 10Z)*
50	95	+1.64	66
49	82.5	+0.93	59
48	65	+0.39	54
47	42.5	-0.19	48
46	22.5	-0.76	42
45	10	-1.28	37
44	2.5	-1.96	30

It will be seen that differences of one unit in the original scores are represented on the *T* scale as varying between 5 and 7.

Problems

A. (a) What T score will correspond to the 5th percentile,
(b) the 16th percentile,
(c) the 99th percentile,
(d) the 50th percentile?

B. What is the score corresponding to the median in Table 4:1?

Answers

A. (a) 34, (b) 40, (c) 73, (d) 50.
B. 47.3.

4. *IQ*'s

IQ's are intelligence quotients and are used in reporting intelligence test scores. Nowadays nearly all I.Q's are deviation I.Q's. Raw scores are converted to a scale with a given mean and standard deviation. In the case of Wechsler Scales the mean is 100 and the standard deviation is 15. If raw scores are normally distributed one can convert them into I.Q's with a mean of 100 and standard deviation of 15 by:

(1) Converting to Z scores.
(2) Multiplying the Z by 15.
(3) Adding or subtracting the result from 100.

A formula which does this is:

$$I.Q.(M,100;\sigma,15)=\frac{(X-M)}{\sigma}+100 \qquad (4:4)$$

To convert I.Q's into percentiles it is necessary to convert them to Z scores and then find the percentile by reference to tables for the normal curve.

Problems

A. Given raw scores with a mean of 80 and a standard deviation of 20, convert the following into I.Q's on a scale with a mean of 100 and a standard deviation of 15.

(a) 20; (b) 90; (c) 100; (d) 67.

A. Now convert each one to an I.Q. on a scale with a mean of 100 and a standard deviation of 24.

B. Convert the following I.Q's to percentiles. The test has a mean of 100 and a standard deviation of 15.

(a) 100; (b) 90; (c) 130; (d) 115;
(e) 110; (f) 70; (g) 85.

Answers

A. (a) 55; (b) 107.5 (c) 115; (d) 90.

B. (a) 28; (b) 112 (c) 124; (d) 84.

C. (a) 50th; (b) 25th; (c) 98th; (d) 84th; (e) 75th; (f) 2nd; (g) 16th.

5. *Sten Scores*

The main tests on which sten scores are used are the personality questionnaires prepared by R. B. Cattell and his associates. Cattell defines stens as:

'Units in a standard ten scale, in which ten score points are used to cover the population range in fixed and equal standard deviation intervals, extending from $2^1/_2$ standard deviations above the mean (sten 10). The mean is fixed at 5.5 stens.' (Cattell, 1965, *The Scientific Study of Personality,* London: Pelican, p.374.

The units on the sten scale are thus half a standard deviation in width, which is fairly coarse grouping compared with *T* scores, (one tenth of a standard deviation), and Wechsler I.Q's, one fifteenth of a standard deviation. Each sten covers a range of percentiles as shown in Table 4:3 which also shows the range of *T* Scores and Wechsler type I.Q's corresponding to stens. This last information is included because Factor B of the 16 Personality Factor Questionnaire is a measure of intelligence.

TABLE 4:3 THE RELATIONSHIP BETWEEN STEN SCORES AND OTHER SCORES

	Upper limit of sten	*Percentiles at upper limits*	*T Scores of range covered by the Sten*	*I.Q's*
10				
9	9.5	97.72	70	130
8	8.5	93.32	65	122.5
7	7.5	84.13	60	115
6	6.5	69.15	55	107.5
5	5.5	50	50	100
4	4.5	30.85	45	92.5
3	3.5	15.85	40	85
2	2.5	6.68	35	77.5
1	1.5	2.28	30	70

(No upper limit is given for Sten 10, and no lower limit for Sten 1, as these would not be sensible.)

Problems

An individual obtains the following test results on a series of intelligence tests.

(a)	Test A (*M* 100, σ 15)		145
(b)	Test B		84th percentile
(c)	Test C	*T* Score	61
(d)	Test D (*M* 100, σ 24)		148
(e)	Test E	Sten Score	9

Convert these scores into:

A. Sten Scores

B. Percentiles.

C. I.Q's (*M* 100, σ 15).

D. *T* Scores.

E. Scores on a test with (*M* 500, σ 50).

Answers

A. (a) 10; (b) 7; (c) 8; (d) 10.

B. (a) 99.9; (c) 86.4; (d) 98; (e) 93-98 percentile.

C. (b) 115; (c) 117; (d) 130; (e) 122-130.

D. (a) 80; (b) 60; (d) 70; (e) 65-70.

E. (a) 650; (b) 550; (c) 555; (d) 600; (e) 575-600.

6. *A Table showing the Relationships between Percentiles, Z Scores, Wechsler I.Q's and T Scores*

TABLE 4:4 THE RELATIONSHIPS BETWEEN PERCENTILES, Z SCORES, WECHSLER I.Q'S AND *T* SCORES

Percentile	*Z Score*	*I.Q.*	*T Score*
1st	- 2.33	65	27
5th	- 1.64	75	34
10th	- 1.28	81	37
15th	- 1.04	84	40
20th	- 0.84	87	42
25th	- 0.67	90	43
30th	- 0.52	92	45
35th	- 0.39	94	46
40th	- 0.25	96	48
45th	- 0.13	98	49
50th	0.00	100	50
55th	+ 0.13	102	51
60th	+ 0.25	104	52
65th	+ 0.39	106	54
70th	+ 0.52	108	55
75th	+ 0.67	110	57
80th	+ 0.84	113	58
85th	+ 1.04	116	60
90th	+ 1.28	119	63
95th	+ 1.64	125	66
99th	+ 2.33	135	73

7. *A general Formula for converting Scores on a Scale with given Mean and Standard Deviation into Scores on a Scale with different Mean and Standard Deviation*

In the previous sections of this chapter the conversion of scores from one scale to another has usually been through the use of *Z* scores. This has been done to emphasize the logic of the procedure. The steps have been:

(1) find the *Z* score on Scale 1

$$\left(\frac{X_1 - M_1}{\sigma_1} = Z\right)$$

and

(2) convert the *Z* score to a score on Scale 2 by (a) multiplying the *Z* score by the standard deviation of
Scale 2 and (b) adding the mean

$$\left(X_2 = Z\sigma_2 + M_2\right)$$

But as the *Z* score on both Scales will be the same by definition:

$$X_2 = \left(\frac{\sigma_2}{\sigma_1}\right)X_1 - \left[\left(\frac{\sigma_2}{\sigma_1}\right)M_1 - M_2\right] \qquad (4:5)$$

Proof

(1) $\frac{X_2 - M_2}{\sigma_2} = \frac{X_1 - M_1}{\sigma_1}$, by definition.

(2) Multiplying both sides by σ_2 gives:

$$X_2 - M_2 = \left(\frac{\sigma_2}{\sigma_1}\right)(X_1 - M_1) = \left(\frac{\sigma_2}{\sigma_1}\right)X_1 - \left(\frac{\sigma_2}{\sigma_1}\right)M_1$$

(3) Adding M_2 to both sides gives

$$X_2 = \left(\frac{\sigma_2}{\sigma_1}\right)X_1 - \left(\frac{\sigma_2}{\sigma_1}\right)M_1 + M_2$$

(4) Thus:

$$X_2 + \left(\frac{\sigma_2}{\sigma_1}\right)X_1 - \left\lfloor \left(\frac{\sigma_2}{\sigma_1}\right)M_1 - M_2 \right\rfloor$$

For many purposes this formula is easier to use than the procedure using *Z* scores. As an example of its use suppose that it is desired to convert a WAIS I.Q. of 90 into a *T* Score. Substituting the appropriate value in the formula gives:

$$T\ \text{Score} = \left(\frac{10}{15}\right)90 - \left\lfloor \left(\frac{10}{15}\right)100 - 50 \right\rfloor = 43.$$

In any situation where there are a large number of scores to convert from one scale to another, the fact that:

$$\frac{\sigma_2}{\sigma_1}(M_1 - M_2)$$

will be a constant will ease the computational burden. If only one score is to be converted, the following variant of (4:5) should be used

$$X_2 = \frac{\sigma_2}{\sigma_1}(X_1 - M_1) + M_2 \qquad (4:6)$$

Chapter Five

Correlation and regression

1. *Introduction*

Up to this point we have been concerned with single variables. The present chapter will discuss relationships between two variables, and the measurement of this relationship by means of the product-moment correlation coefficient. This coefficient varies in value between 0 and 1.0. It can be positive or negative in sign. If scores on one variable rise as scores on the other one rise then the correlation is positive, while if scores on one variable fall as the other scores rise the correlation is negative. For example height and weight are positively correlated because taller people tend to be heavier than smaller people, while mental speed and age are negatively correlated in adults, as mental speed drops with increasing age. If there is no relationship at all between two variables the correlation is zero. If the relationship is perfect, i.e. if there is complete correspondence between the two variables, the correlation will be 1.0. (Complete correspondence in this case is indicated by individuals obtaining exactly the same *Z* score on both variables).

An important point to bear in mind is that the product-moment correlation coefficient measures the strength of a linear relationship between two variables. If the relationship between two variables is not linear, then the correlation coefficient will be of little use. A linear relationship exists when the graph showing the relationships between two variables is a straight line, or near enough to a straight line, for a straight line to be a reasonable approximation to it. Figure 5.1 shows some linear and non-linear relationships.

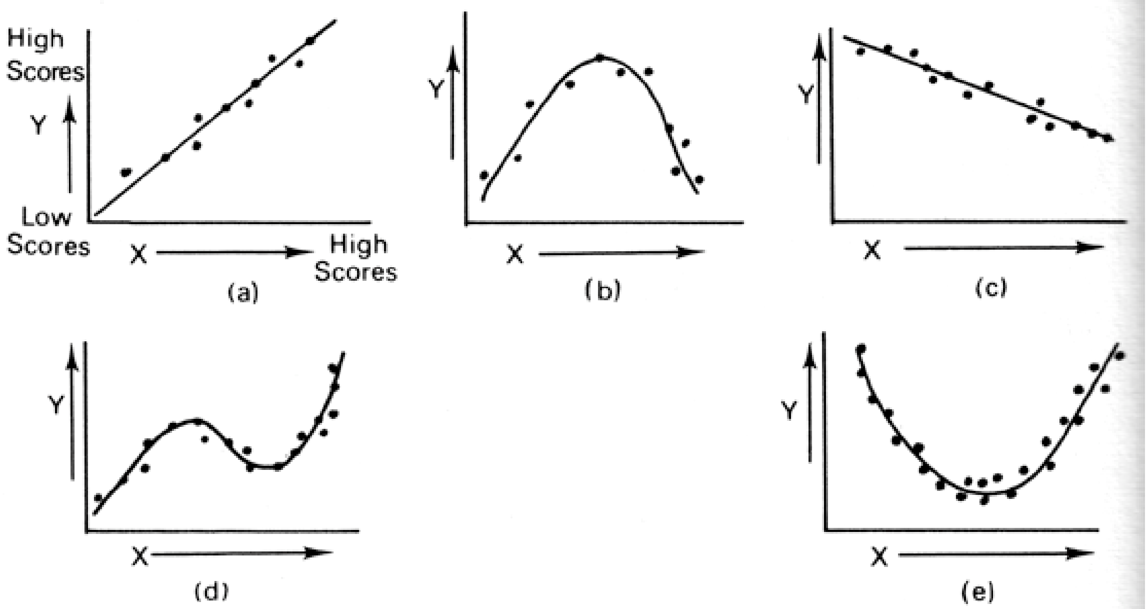


Figure 5.1 Some linear and non-linear relationships between two variables *X* and *Y*

Problems

A. Which of the relationships in Figure 5.1 can be adequately described by a product-moment correlation coefficient?

B. Which of the relationships represents a negative correlation?

C. Which relationships are curvilinear?

Answers

A. a; c.

B. c.

C. b; d; e.

2. *The Basic Formulae for the Product-Moment Correlation Coefficient.*

Suppose that we have two variables X and Y, the product moment correlation coefficient between them is symbolised r_{xy} and its formula is:

$$r_{xy} = \frac{\sum Z_X Z_Y}{N} \qquad (5:1)$$

From the formula it can be seen that the correlation coefficient is the mean of the products of the *Z* scores. To obtain the coefficient by this formula we need to take the two *Z* scores obtained by an individual, i.e. his *Z* score on *X* and his *Z* score on *Y*, and multiply them together. This is repeated for all individuals and the products so obtained are summed. This sum is then divided by *N*, and the result is the value of r_{xy}.

As an example suppose that seven individuals complete tests *X* and *Y* and obtain the following scores:

	Scores		
Individuals	*Test X*	*Test Y*	
A	1	12	
B	2	14	Mean X = 4
C	3	10	σ_x= 2.0
D	4	6	
E	5	8	Mean Y = 8.0
F	6	2	σ_y= 4.0
G	7	4	

Converting these scores into *Z* scores and finding the products of each pair of *Z* scores gives the following:

Individuals	Z_X	Z_Y	$Z_X Z_Y$
A	-1.5	+1.0	-1.5
B	-1.0	+1.5	-1.5
C	-0.5	+0.5	-0.25
D	0	-0.5	0
E	+0.5	0	0
F	+1.0	-1.5	-1.5
G	+1.5	-1.0	-1.5

$\sum Z_X Z_Y = -6.25$ $N = 7$ $r_{xy} = -0.89$

This version of the formula for the product moment correlation coefficient is not the one most commonly found in basic text books, but it will be an extremely useful one for our purposes.

By simple manipulation this formula can be converted into a more common one, which will also be useful later.

$$r_{xy} = \frac{\sum xy}{N\sigma_x \sigma_y} \tag{5:2}$$

Proof

(1) $$r_{xy} = \frac{\sum Z_x Z_y}{N}$$

(2) $$Z_x = \frac{x}{\sigma_x}; \quad and \quad Z_Y = \frac{y}{\sigma_y}.$$

(3) So $r_{xy} = \dfrac{\sum\left(\dfrac{x}{\sigma_x}\cdot\dfrac{y}{\sigma_y}\right)}{N}$

(4) Multiplying numerator and denominator by σ_x. σ_y gives

$$\frac{\sum xy}{N\sigma_x\sigma_y}$$

Neither of these formulae is very convenient for computational purposes so further operations would be necessary to derive easily computable formulae, but in the forms given they will be ideal for our purposes. At this stage it should be noted that by multiplying both sides of 5:2 by $.\sigma_X$ $.\sigma_Y$ we obtain:

$$\frac{\sum xy}{N\sigma_x\sigma_y} \qquad (5:3)$$

The term on the right is called the covariance of *X* and *Y*. Covariances and formula (5:3) will be used frequently in later sections.

3. *The Scatter Diagram*

A scatter diagram is a graphical device for showing the distribution of scores on two variables. The diagram is constructed by taking all subjects with a given score, X_1, on one variable and plotting the distribution of *Y* scores for these individuals. This will be the first column of the scatter diagram. Next the *Y* scores for all obtaining score X_2 are plotted, forming the second column and so on. As an example suppose that the following scores are obtained on two tests.

	Tests			*Tests*	
Individuals	*X*	*Y*	*Individuals*	*X*	*Y*
A	1	1	F	3	4
B	1	2	G	3	3
C	2	3	H	4	4
D	2	2	I	4	5
E	2	3	J	5	4

These scores can be plotted on a scatter diagram as shown in Figure 5.2.

Y

Scores

6							
5					1		
4				1	1	1	
3			2	1			
2		1	1				
1		1					
0							
	0	1	2	3	4	5	6

X Scores

Figure 5.22 A scatter diagram of the data in table 5.3

Of the two subjects scoring 1 on test *X*, one obtained a score of 1 on test *Y*, and the other a score of 2. Of the three obtaining a score of 2 on test *X*, two scored 3 on *Y* and one scores 2, and so on. The scatter diagram can of course be read both ways. It is easy to see what *X* scores were obtained by people scoring 3 on *Y* for example.

Usually there is a much larger number of scores involved in a scatter diagram. If we draw a scatter diagram without grid lines and represent individuals by dots, we might obtain something by dots, we might obtain something like that shown in Figure 5.3.

Figure 5.3 Another scatter diagram

In this scatter diagram there is a tendency for those scoring high on X to also score high on Y. If the correlation was 1.0 there would be no scatter of scores within columns, and the points would all fall on a straight line. When the correlation is zero the scores will fall in a circular pattern, (if the variables are normally distributed). As the correlation rises from 0 to 1.0, the shape of the scatter becomes more elongated and ellipsoid until it becomes a single straight line. Figure 5.4 shows these changes visually.

Figure 5.4 Changes in the outline of the scatter plot as the correlation between two variables increases

Changes in the range of scores on one variable will (a) affect the range of scores on the other, and (b) the size of the correlation coefficient obtained. Suppose that the relationship between *X* and *Y* for the full range of scores is as shown in Figure 5.5. This would represent a reasonably high degree of correlation. If, however, we had a sample of subjects whose range of ability on *X* fell in the range *A* to *B*, this would curtail the range of *Y* scores to the range *C* to *D*. The shape of the scatter diagram obtained for this group would be as depicted in the smaller figure on the right in Figure 5.5.

Figure 5.5 The effects of restriction of range on the shape of the scatter plot

This smaller scatter diagram is more like the pattern of zero correlation, than is the scatter diagram for the full range, and indeed a restriction in range generally reduces the value of the correlation coefficient, (see Chapter 12, Section 4).

Problem

An investigator interested in the general relationship between creativity and intelligence, after finding a low relationship between these variables in a sample of university students, concludes that creativity is largely independent of intelligence. Should he have done so?

4. *Correlation and Predicition:*

(1) Guessing

The main use of correlation coefficients is prediction. The existence of a significant correlation coefficient means that X scores can be predicted from *Y* scores with better than chance accuracy. If there were no correlation between *X* and *Y* then knowing the individual's *X* score would tell us nothing about the likely *Y* score. If we know nothing about an individual's *Y* score and we have to guess it, then our best bet is that the score obtained will be the mode. If we know what outcomes are possible, and do not know which outcome will occur, our best bet is that the most common outcome will occur. This will lead to fewer mistakes in the long run than any other bet. If you know that someone has a set of cards consisting of two hearts and 11 spades in their hand, and someone chooses one at random, and you have to guess what it is, your best bet is that it is a spade. Following exactly the same principle the mode is the best bet in the case of test scores. With normal distributions the mode is the same as the mean. So the best strategy in attempting to predict Y from X or *X* from *Y* in the absence of any correlation between them is to choose the mode which in the case of test scores will usually be the mean. By choosing the mode we will be absolutely right more frequently than by choosing any other value.

However, being absolutely right is not the only criterion we might choose. In predicting test scores we might be more concerned with the average distance of our predictions from the true value. We might decide that we want our *average error* to be as near to zero as possible. Suppose in the absence of other information we guess the mean as the most likely score, i.e. for each individual and then compute the differences between the obtained score *X* and the predicted score, the mean, we will have a distribution of *(X - M)*'s summing across individuals, $\Sigma(X - M)$ is obtained, and this we have seen earlier (2:3) is equal to 0. So if we choose the mean the average error of our predictions will be zero. It is possible to show that this will be true of no other value. Suppose that a different value *D* is chosen, where *D* is a score other than the mean.

(1) D = Score other than the mean.

(2) $D = M - A$ (where A is a positive or negative number).

(3) $X - D = X - (M - A).$
(4) $X - (M - A) = X - M + A.$

(5) Therefore $\Sigma(X - D) = \Sigma\ (X - M + A).$

(6) So $\Sigma(X - D) = \Sigma\ X - NM + NA.$

(7) But $M = \Sigma\ X/N,$ so $\Sigma X = NM.$

(8) So (6) becomes $NM - NM + NA.$

(9) So $\Sigma\ (X - D) = NA.$

(10) And $\sum \frac{(X-D)}{N} = \frac{NA}{N} = A.$

The value of *A* differs from zero, therefore, choosing a point other than the mean leads to a greater average error than would have ensued from the choice of the mean. Thus if our interest is in obtaining the smallest average deviation between guessed and actual score we choose the mean.

To summarize this section:

(1) If it is important to be absolutely right, guess the mode.
(2) If it is desirable that average error in prediction should be zero, then guess the mean.

If we are interested in the smallest absolute average deviation, the best measure of central tendency to use, and the best guess to make, would be the median. The sum of absolute deviations, i.e. deviations disregarding their sign, is smaller when the median is used than when any other point is chosen.

Fortunately most test scores are normally distributed and thus the mean = the mode = the median. So there is no problem in deciding which to use as the best bet.

5. *Correlation and prediction: (2) Linear Regression*

Suppose that two tests have been administered to a group of subjects and the following scores obtained.

	Tests			*Tests*	
Subject	*X*	*Y*	*Subject*	*X*	*Y*
A	0	0	E	4	20
B	1	5	F	5	25
C	2	10	G	6	30
D	3	15			

If these scores are plotted on a graph with X as the horizontal axis and Y as the vertical one, it will be seen that a straight line fits the points exactly. This is shown in Figure 5.6.

Figure 5.6

We can also construct an equation $Y = bX$ for predicting Y scores from X scores. In this case $Y - 5X$, each Y value is five times the corresponding X value.

The value b tells us the slope of the line. By definition the slope of a line is given by taking two Y values, say Y_1 and Y_2 and their corresponding X values, X_1 and X_2 , and finding the value of:

$$\frac{Y_2 - Y_1}{X_2 - X_1}$$

This slope is equal to the ratio of the change in the Y variable to the change in the X variable. If Y goes down as X goes up, slope is negative, while if both rise together, slope is positive.

The next concept to be introduced is the intercept. The Y intercept is the point where the line crosses the Y axis. Consider the following sets of scores.

	Tests			*Tests*	
Subject	*X*	*Y*	*Subject*	*X*	*Y*
A	0	5	E	4	25
B	1	10	F	5	30
C	2	15	G	6	35
D	3	20			

If we plot these again the relationship is linear as in Figure 5.7.

Figure 5.7

But this time the line cuts through the *Y* axis at the value of 5. Therefore, a simple formula of the type $Y = bX$ will no longer suffice. The formula has to be modified by taking the intercept into account. The symbol for an intercept is *'a'*.

$$Y = bX + a \qquad (5.4)$$

This is the general formula for a linear relationship.

For the data presented above *b* can be found to be 5 and *'a'* can be seen to be 5, thus $Y = 5X + 5.$

In psychology, data, seldom, if ever, falls exactly on a straight line. A group of individuals obtaining a given *X* score will not all get the same *Y* score. Even if we compute the means for each group with a given *Y* score, the means are not likely to lie on a straight line. For example, let us plot the following data:

	Tests			*Tests*	
Individuals	*X*	*Y*	*Individuals*	*X*	*Y*
A	1	2	G	3	4
B	1	3	H	3	5

C	1	4	I	3	6
D	2	3	J	4	5
E	2	5	K	4	6
F	2	5	L	4	8

It can be seen in Figure 5.8 that although the points do not lie upon a straight line it is obvious that a linear prediction rule might have some value here.

Figure 5.8

The problem is how do we find a straight line to fit the data, when the points do not lie in a straight line. We need some criterion by which to choose amongst possible straight lines which might be fitted to the data. The criterion used is the least squares criterion.

The line of best fit is defined to be that line which minimizes the squared deviations between predicted and obtained scores. A line so chosen is known as a regression line.

If we decide that we want a linear equation for predicting *Z* scores on test *Y,* symbolized $\left(\hat{Z}_y\right)$, from *Z* scores on test *X, (Z_x),* then will need a formula of the following type:

$$\hat{Z}_y = bZ_x + a$$

$\hat{Z}_y$ is used rather than Z_y to indicate that it is an estimate of Z_y which ill differ from Z_y by the quantity $\hat{Z}_y - Z_y$. The least squares principle states that we must choose the values in the equation to make $\sum\left(\hat{Z}_y - Z_y\right)^2 / N$, as small as possible. It can be demonstrated that for this to be true:

$$a = 0 \qquad (5.5)$$

Proof

(1) $\hat{Z}_Y = bZ_X + a$

(2) So $Z_Y - \hat{Z}_Y = \left(Z_Y - bZ_X\right) - a$

(3) and $\left(Z_Y - \hat{Z}_Y\right)^2 = \left(Z_Y - bZ_X\right)^2 + a^2 - 2a\left(Z_Y - bZ_X\right)$

(4) so $\sum\left(Z_Y - \hat{Z}_Y\right)^2 = \sum\left(Z_Y - bZ_X\right)^2 + Na^2 - 2a\sum\left(Z_Y - bZ_X\right)$

(5) and $\dfrac{\sum\left(Z_Y - \hat{Z}_Y\right)^2}{N} = \dfrac{\sum\left(Z_Y - bZ_X\right)^2}{N} + a^2 - 2a\left(\dfrac{\sum Z_Y}{N} - b\dfrac{\sum Z_X}{N}\right)$

(6) as $\dfrac{\sum Z_Y}{N} = \overline{Z}_Y \; and \; \dfrac{\sum Z_X}{N} = \overline{Z}_X$

these both = 0, and therefore $2a\left(\dfrac{\sum Z_Y}{N} - b\dfrac{\sum Z_X}{N}\right) = 0$

(7) thus $\dfrac{\sum\left(Z_Y - \hat{Z}_Y\right)^2}{N} = \dfrac{\sum\left(Z_Y - bZ_X\right)^2}{N} + a^2$

(8) as a^2 must be positive, (all squared numbers are), for $\frac{\sum\left(Z_Y - \hat{Z}_Y\right)^2}{N}$ to be at its lowest '*a*' must equal zero

Having demonstrated that '*a*' must be zero, we can simplify the equation thus:

$$\hat{Z} = bZ_X \qquad (5:6)$$

It can also be shown that bZ_X must equal $r_{xy}Z_X$ if the least squares criterion is to be met.

Proof

(1) $Z_Y - \hat{Z}_Y = Z_Y - bZ_X$

(2) $\left(Z_Y - \hat{Z}_Y\right)^2 = Z_Y^2 + b^2 Z_X^2 - 2bZ_X Z_Y$

(3) $\sum\left(Z_Y - \hat{Z}_Y\right)^2 = \sum Z_Y^2 + b^2 Z_X^2 - 2b\sum Z_X Z_Y$

(4) $\frac{\sum\left(Z_Y - \hat{Z}_Y\right)^2}{N} = \frac{\sum Z_Y^2}{N} + b^2 \frac{\sum Z_X^2}{N} - 2b\frac{\sum Z_X Z_Y}{N}$

(5) $\frac{\sum Z_Y^2}{N} = \sigma_Z^2 = 1; \frac{\sum Z_X^2}{N} = \sigma_Z^2 = 1;$

and $\frac{\sum Z_X Z_Y}{N} = r_{xy}$

So we obtain

$$\frac{\sum\left(Z_Y - \hat{Z}_Y\right)^2}{N} = 1 + b^2 - 2br_{xy}$$

(6) It will now be shown that if *b* is any value other than r_{xy} then $1 + b^2 - 2br_{xy}$ will be larger in value than if *b* is equal to r_{xy}

(a) if $b = r_{xy}$ then (5) becomes $1 + r^2_{xy} - 2r^2_{xy} = 1 - r^2_{xy}$

(b) if *b* was other than r_{xy}, say $r_{xy} - C$ then (5) becomes $1 + (r_{xy} - C)^2 - 2(r_{xy} - C)r_{xy}$

(c) this equals: $1 + r^2_{xy} + C^2 - 2Cr_{xy} - 2r^2_{xy} + 2Cr_{xy}$ $= 1 - r^2_{xy} + C^2$. Which is 6(a) + C^2

(d) C^2 must be positive as it is a squared value, so 6(c) must be larger than 6(a). Therfore, r_{xy} is the value which gives the smallest value of $\sum\left(Z_Y - \hat{Z}_Y\right)^2$

Equation (5:6) is called the linear regression equation for predicting $\hat{Z}_Y$ from Z_X. For raw scores the linear regression equation will be:

$$\hat{Y} = r_{xy}\frac{\sigma_y}{\sigma_x}\left(X - M_x\right) + M_y \qquad (5:7)$$

Problems

A. What will be the value of $\hat{Z}_Y$ from Z_X. For raw scores the linear regression equation will be:

$$\hat{Y} = r_{xy}\frac{\sigma_y}{\sigma_x}\left(X - M_x\right) + M_y \qquad (5:7)$$

Problems

A. What will be the value of $\hat{Z}_Y$ when Z_X equals $\overline{Z}_X$?
B. What will be the value of $\hat{X}$ when *Y* equals M_y?

Answers

A. $\hat{Z}_Y = r_{xy}Z_x$; therefore when $Z_X = Z_X = 0$.
$\hat{Z}_Y = 0 \times r_{xy} = 0 = \overline{Z}_Y$.

B. $\hat{X} = r_{xy}\frac{\sigma_x}{\sigma_y}\left(Y - M_y\right) + M_x$; therefore when $Y = M_y$,

$$X = r_{xy}\frac{\sigma_x}{\sigma_y}\left(M_y - M_y\right) + M_x = M_x$$

In both cases, when the predictor variable assumes its mean value, the predicted value becomes the mean of the criterion variable.

Thus a regression line passes through the point of intersection of M_x and M_y. It is also true that:-

$$\hat{M}_{y=}M_y \qquad (5:8)$$

Proof

(1) $\hat{M}_y = \frac{\sum \hat{Y}}{N}$

(2) $$\frac{\sum \hat{Y}}{N} = \frac{\sum\left(M_y + r_{xy}\left(\sigma_y / \sigma_x\right)\left[X - M_x\right]\right)}{}$$

(This is obtained by use of Formula (5:7)

(3) Therefore;

$$\frac{\sum \hat{Y}}{N} = \frac{NM_y + r_{xy}\left(\sigma_y / \sigma_x\right)\left[\sum X - NM_x\right]}{N}$$

$$= M_y = r_{xy}\frac{\sigma_y}{\sigma_x}\left[M_x - M_x\right]$$

(4) Therefore: $\hat{M}_y = M_y$

We can now prove Formula (5:7)

Proof

(1) $\hat{Z}_Y = \dfrac{\hat{Y} - \hat{M}_y}{\sigma_y} = \dfrac{\hat{Y} - M_y}{\sigma_y}$

(2) and $\hat{Z}_Y = r_{xy} Z_X$

(3) but $Z_X = \dfrac{X - M_x}{\sigma_x}$

(4) So $\hat{Z}_Y = \frac{\hat{Y} - M_y}{\sigma_y} = r_{xy} \frac{(X - M_x)}{\sigma_x}$

(5) Multiplying the last two terms of (4) by σ_y gives

$$\hat{Y} - M_y = \sigma_y r_{xy} \frac{(X - M_x)}{\sigma_x}$$

(6) A little rearrangement gives:

$$\hat{Y} - M_y = r_{xy} \frac{\sigma_y}{\sigma_x}(X - M_x)$$

(7) Adding M_y to both sides we obtain:

$$\hat{Y} = r_{xy} \frac{\sigma_y}{\sigma_x}(X - M_x) + M_y$$

This, as stated above, is the raw score linear regression equation for predicting *Y* from *X*.

Problems

Given two tests *X* and *Y* with $M_x = 50$; $\sigma_x = 10$, and $M_y = 100$, $\sigma_y = 20$, and $r_{xy} = +\ 0.80$:

A. Find the predicted $\hat{Z}_Y$ for someone who scores 30 on test *X*.

B. Find the predicted raw score $(\hat{Y})$ for someone who scores 90 on text *X*.

Answers

A. $\hat{Z}_Y = r_{xy} Z_X, \textit{ and } Z_X = \dfrac{30-50}{10} = -2.0$

So $r_{xy}\, Z_X = 0.80 \times (-0.20) = -1.60$

So $\hat{Z}_Y = -1.60$

B. $\hat{Y} = r_{xy} \dfrac{\sigma_y}{\sigma_x}(X - M_x) + M_y$ so

$$\hat{Y} = 0.80 \frac{20}{10}(90-50) + 100 = 164$$

The slope of the regression line of the *Y* scores on the *X* scores, i.e. the best fit line for predicting *Y* from *X*, has been shown to be $r_{xy}(\sigma_y / \sigma_x)$. If in (5:7) we had been concerned with predicting *X* from *Y* instead of *Y* from *X* we would have found that instead of $r_{xy}(\sigma_y / \sigma_x)$ we would have obtained $r_{xy}(\sigma_x / \sigma_y)$.

This would be the slope of the regression line for predicting *X* from *Y*. In both cases the slope is the product of the correlation coefficient and the ratio of the standard deviations, and in both cases the standard deviation of the predicted variable is the numerator of the ratio. The moral of this tale is that except in the case where $r_{xy} = 1.0$ there will be two regression lines in the scatter diagram one with slope $b_{y.x}$ and one with slope $b_{x.y}$. The subscript y.x means *Y* predicted from *X*, and x.y means *X* predicted from *Y*. The slopes of the regression lines as correlation increases are shown in Figure 5.9, where it can be seen that as r_{xy} increases, the regression lines get closer together until at a correlation of 1.0 they become one line.

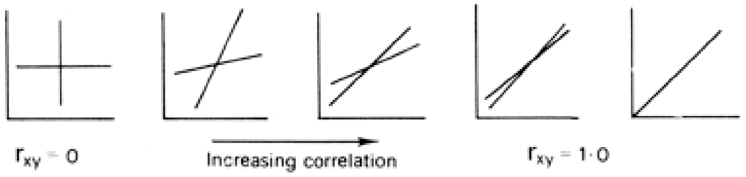


Figure 5.9 Changes in the relationships of the regression lines as correlation increases

Figure 5.9 Changes in the relationships of the regression lines as correlation increases

The interpretation of correlation coefficients

1. *The Standard Error of Estimate*

In the last chapter it was shown that the slope of the regression line in Z score terms is r_{xy}. During the proof that if the least squares principle is adopted the slope must be r_{xy}, it was shown that:

$$\frac{\sum\left(Z_Y - \hat{Z}_Y\right)^2}{N} = 1 - r_{xy}^2 \qquad (6:1)$$

(This was demonstrated in steps 5 and 6 of the proof that *b* must equal r_{xy}. See page 47).

Recalling that the formula for the variance is $\sum\frac{(X-M)^2}{N}$ it can be seen that the left hand side of (6:1) has some similarities to the variance. The difference is that in (6:1) the deviations are not from the mean but from the predicted score, which will of course fall on the regression line. Hence the deviations in (6:1) are deviations of scores about the regression line. Obtained *Y* scores will be normally distributed about the regression line with a mean of $r_{xy}Z_x$ and a variance of $\frac{\sum\left(Z_Y - \hat{Z}_Y\right)^2}{N} or\, 1 - r_{xy}^2$. The standard deviation of the distribution will be $\sqrt{1 - r_{xy}^2}$. This value is called the standard error of estimate, and is the standard deviation of errors of prediction about the regression line.

Standard error of estimate = σ_{est}. (6:2)

= (a) (in Z scores) $\sqrt{1 - r_{xy}^2}$
= (b) (in raw scores) $\sigma \sqrt{1 - r_{xy}^2}$

The standard deviation involved in (6:2(b)) is the standard deviation of the scores being predicted.

The standard error of estimate is to be regarded just like an ordinary standard deviation. Provided that the assumptions of (a) normal distributions of scores in the rows and columns of the scatter diagram, and (b) of equal standard deviations in columns, and (c) of equal standard deviations in rows are not seriously violated it can be used for assessing the likelihood of deviations of a given magnitude from the regression line. (Note that assumptions (b) and (c) are separate; the standard deviations of columns do not have to be the same as the standard deviations of rows.) The similarity of standard deviations and standard error of estimate is emphasized in Table 6:1, which can be used to solve the problems following it.

TABLE 6:1
SIMILARITY OF STANDARD DEVIATION AND
STANDARD ERROR OF ESTIMATE

$\frac{x}{\sigma} = Z$	*Proportion of cases between M and Z*	$\frac{Y - \hat{Y}}{\sigma\sqrt{1 - r_{xy}^2}}$	*Proportion of cases lying between Y and regression line*
.00	.0000	.00	.0000
.10	.0398	.10	.0398
.20	.0793	.20	.0793
.50	.1915	.50	.1015
1.00	.3413	1.00	.3413
2.00	.4772	2.00	.4772
3.00	.49865	3.00	.49865

Problems

A. A subject obtains a *T* score of 70 on a test of intelligence, which is known to have a correlation of +.60 with examination marks. The examination result shows the subject to be at the 50th percentile. Has he done worse than was expected?

B. If two tests *X* and *Y* correlate .8 with each other and M_x is 100, and σ_x is 15 while M_y is 50, and σ_y is 10, what proportion of people with an average score on *X* would you expect to score between 44 and 56 on *Y*?

Answers

A. Using *Z* scores:-

(a) $\hat{Z}_Y = r_{xy} Z_X = .60 \times 2 = +1.20$

(b) The score on *Y* is at the 50th percentile so $Z_Y = 0$

(c) The standard error of estimate = $\sqrt{1 - r_{xy}^2} = \sqrt{1 - .60^2} = .80$

(d) $\left(Z_Y - \hat{Z}_Y\right)\sigma_{est.} = \frac{0 - 1.20}{.80} = -1.50$

So the subject's exam mark is 1.5 standard errors of estimate below the expected mark. Reference to Table 6:1 shows that if it had been 1.0 standard error below 84.13 per cent would have done better so we can say that over 84.13 per cent of students of this subject's intellectual level would have been expected to do better in the examination.

B. Using raw scores

(1) $\hat{Y} = r_{xy} \frac{\sigma_y}{\sigma_x} \left(X - M_x\right) + M_y$

(2) $= .80 \frac{10}{15} \left(X - M_x\right) + 50 = 50$

$\sigma_{est.} = \sigma_y \sqrt{1 - r_{xy}^2} = 10\sqrt{1 - .80^2} = 6.0$

(3) Treating 56 as an obtained score *Y* we can find the

ratio $\frac{Y - \hat{Y}}{\sigma_{est.}} = \frac{56 - 50}{6} = +1.0$

So consulting Table 6:1, 34.13 per cent of people scoring at the mean on test *X* would be expected to score between 50 and 56 on test *Y*.

(4) Repeating these operations for *Y* = 44 we find

$\frac{44-50}{6}=-1.0.$ So 34.13 per cent would be expected to score between 44 and 50 on *Y*.

(5) Adding these together gives 68.26 per cent which is the percentage of those scoring at the mean on *X* who would be expected to score between 44 and 56 on *Y*.

Let us consider what happens to the difference between predicted scores and predictor scores as the correlation coefficient changes in value. If the correlation is zero:

(1) the best estimate of the predicted *Z* score is zero.

(2) the standard error of estimate will have the same value as the standard deviation.

Both of these statements can be easily verified by substituting zero for r_{xy} in the appropriate formulae.

If the correlation is 1.0:

(1) the best estimate of the predicted score is that score which has the same *Z* score value as the predictor score.

(2) the standard error of estimate will be zero. All scores will fall on the regression line and there will be no scatter of scores around it.

Again it can be easily verified that these statements are correct. Table 6:2 shows changes in the accuracy of prediction as r_{xy} changes, assuming a Z_x of +3.0.

TABLE 6:2
ACCURACY OF PREDICTION AS A FUNCTION
OF r_{xy}

r_{xy}	σ_{est}	$\hat{Z}_Y, (Z_X = 3.0)$
.00	1.00 σ_y	.00
.10	.995 σ_y	.30
.20	.980 σ_y	.60
.30	.954 σ_y	.90
.40	.917 σ_y	1.20
.50	.866 σ_y	1.50
.60	.800 σ_y	1.80
.70	.714 σ_y	2.10
.80	.600 σ_y	2.40
.90	.436 σ_y	2.70
1.00	0	3.00

As r_{xy} increases the size of the standard error of estimate falls, and the difference between the predictor score and the predicted score becomes less, until at r_{xy} = 1.00 $Z_Y = Z_X$ and there is no error of prediction at all.

Sometimes a measure called the index of forecasting efficiency is used. This gives the percentage reduction in σ_{est}. as a function of r_{xy}. Its formula is:

Index of forecasting efficiency = *E*

$$= 100(1\sqrt{1-r_{xy}^2}) \qquad (6:3)$$

Thus if r_{xy} is .80 it can be seen from Table 6:2 that $\sqrt{1-r_{xy}^2} = .600$ so the value of *E* equals 100(1 - .60) = 40. So the standard deviation of error has been reduced by 40 per cent of what it would have been with r_{xy} = 0. It remains nevertheless at 60 percent.

2. *The Variance Accounted for by a Correlation of a Given Size*

It is possible to divide the variance of a predicted variable into two parts:

(1) that accounted for by the predictor variable; and

(2) that not accounted for in this way, the residual variance.

The use of the phrase 'accounted for' is not to be taken in a deterministic way. If two variables are correlated then the proportion of variance in either accounted for by the other will be the same. Putting the above statements as a formula gives

$$\frac{\sum\left(Z_Y-\overline{Z}_Y\right)^2}{N}=\frac{\sum\left(Z_Y-\hat{Z}_Y\right)^2}{N}+\frac{\sum\left(\hat{Z}_Y-\overline{Z}_Y\right)^2}{N} \qquad (6:4)$$

Where

$$\frac{\sum\left(Z_Y-\overline{Z}_Y\right)^2}{N} = \text{total variance}$$

$$\frac{\sum\left(Z_Y-\hat{Z}_Y\right)^2}{N} = \text{residual variance}$$

$$\frac{\sum\left(\hat{Z}_Y-\overline{Z}_Y\right)^2}{N} = \text{variance accounted for}$$

Proof

(1) $Z_Y - \overline{Z}_Y = \left(Z_Y - \hat{Z}_Y\right) + \left(\hat{Z}_Y - \overline{Z}_Y\right)$

(2) So $\left(Z_Y - \overline{Z}_Y\right)^2 = \left(Z_Y - \hat{Z}_Y\right)^2 + \left(\hat{Z}_Y - \overline{Z}_Y\right)^2$
$+2\left(Z_Y - \hat{Z}_Y\right)\left(\hat{Z}_Y - \overline{Z}_Y\right)$

(3) $As\ \overline{Z}_Y = 0, 2\left(Z_Y - \hat{Z}_Y\right)\left(\hat{Z}_Y - \overline{Z}_Y\right) = 2\hat{Z}_Y\left(Z_Y - \hat{Z}_Y\right)$

(4) $\sum\left(\hat{Z}_Y - \overline{Z}_Y\right)^2 = \sum\left(Z_Y - \hat{Z}_Y\right)^2 + \sum\left(\hat{Z}_Y - \overline{Z}_Y\right)^2$
$+2\sum Z_Y\left(Z_Y - \hat{Z}_Y\right)$

(5) However $\hat{Z}_Y = r_{xy}Z_X$ $so\ 2\sum\hat{Z}_Y\left(Z_Y - \overline{Z}_Y\right) = 2\sum r_{xy}Z_X$
$\left(Z_Y - r_{xy}Z_X\right)$

(6) Multiplying this out gives
$2\sum r_{xy}Z_X\left(Z_Y - r_{xy}Z_X\right) = 2r_{xy}\sum Z_X Z_Y - 2r_{xy}^2\sum Z_X^2$

(7) But (a) as $r_{xy} = \frac{\sum Z_X Z_Y}{N}, \sum Z_X Z_Y = Nr_{xy} = Nr_{xy}$ and (b) it has been shown in (3:4) that $\sum Z^2 = N.$ So (6) becomes $2r_{xy}Nr_{xy} - 2r_{xy}^2 N = 0$.

(8) Therefore, from (4)
$\sum\left(Z_Y - \overline{Z}_Y\right)^2 = \sum\left(Z_Y - \hat{Z}_Y\right)^2 + \sum\left(\hat{Z}_Y - \overline{Z}_Y\right)^2$

(9) Dividing by *N* gives the variance:

$$\frac{\sum\left(Z_Y - \overline{Z}_Y\right)^2}{N} = \frac{\sum\left(Z_Y - \hat{Z}_Y\right)^2}{N} + \frac{\sum\left(\hat{Z}_Y - \overline{Z}_Y\right)^2}{N}$$

or (as $\bar{Z}_Y = 0$)

$$\frac{\sum\left(Z_Y - \hat{Z}_Y\right)^2}{N} + \frac{\sum \hat{Z}_Y^2}{N}$$

The residual variance, i.e. that left after the effects of *X* have been removed is $\frac{\sum\left(Z_Y - \hat{Z}_Y\right)^2}{N}$. However from (6:1) this is equal to $1 - r_{xy}^2$. So:-

(a) Total variance of *Z* scores = 1.0.

(b) Residual variance = 1 - r_{xy}^2.

(c) Therefore accounted for variance =

$$1 - \left(1 - r_{xy}^2\right) = r_{xy}^2$$

So the variance accounted for with a total variance of 1 is r_{xy}^2. The proportion of variance accounted for is therefore $\frac{r_{xy}^2}{1} = r_{xy}^2$.

Proportion of variance accounted for = r_{xy}^2 (6:5)

Problems

A. Give a formal proof that $\frac{\sum \hat{Z}_Y^2}{N} = r_{xy}^2$

B. Give a formal proof that the proportion of variance accounted for = r_{xy}^2. Starting with

$$\frac{\text{Accounted for variance}}{\text{Total variance}} = \frac{\frac{1}{N}\sum \hat{Z}_Y^2}{\frac{1}{N}\sum Z_Y^2}$$

Answers

A. (1) $\dfrac{\sum \hat{Z}_Y^2}{N} = \dfrac{\sum \left(r_{xy}^2 Z_X\right)^2}{N}$

(2) $= r_{xy}^2 \dfrac{\sum Z_X^2}{N}$

(3) As $\dfrac{\sum Z^2}{N} = \sigma_Z^2 = 1.0, \dfrac{\sum Z_X^2}{N} = 1.0$

(4) Therefore $\dfrac{\sum \hat{Z}_Y}{N} = r_{xy}^2 (1) = r_{xy}^2$

B. (1) $\dfrac{\frac{1}{N}\sum \hat{Z}_Y^2}{\frac{1}{N}\sum Z_Y^2} = \dfrac{\sum \hat{Z}_Y^2}{\sum Z_Y^2}$ (Numerator and Denominator multiplied by N)

(2) from (6:4) $\sum Z_Y^2 = \sum \hat{Z}_Y^2 + \sum \left(Z_Y - \hat{Z}_Y\right)^2$ and from (A) $\sum \hat{Z}_Y^2 = Nr_{xy}^2$

so $\dfrac{\sum \hat{Z}_Y^2}{\sum Z_Y^2} = \dfrac{Nr_{xy}^2}{Nr_{xy}^2 + \sum \left(Z_Y - \hat{Z}_Y\right)^2}$

(3) But from (6:1) $\sum \left(Z_Y - \hat{Z}_Y\right)^2 = N\left(1 - r_{xy}^2\right)$

$$\text{so}\quad \frac{\sum \hat{Z}_Y^2}{\sum Z_Y^2} = \frac{Nr_{xy}^2}{N\left[r_{xy}^2 + \left(1 r_{xy}^2\right)\right]} = \frac{Nr_{xy}^2}{N(1)}$$

(4) Dividing the right hand term by *N* leaves $\frac{\sum \hat{Z}_Y^2}{\sum Z_Y^2} = r_{xy}^2$

r_{xy}^2 is called the coefficient of determination, and it indicates the proportion of variance in each of two correlated variables which is shared by both. A diagrammatic representation of r_{xy}^2 is given in Figure 6.1.

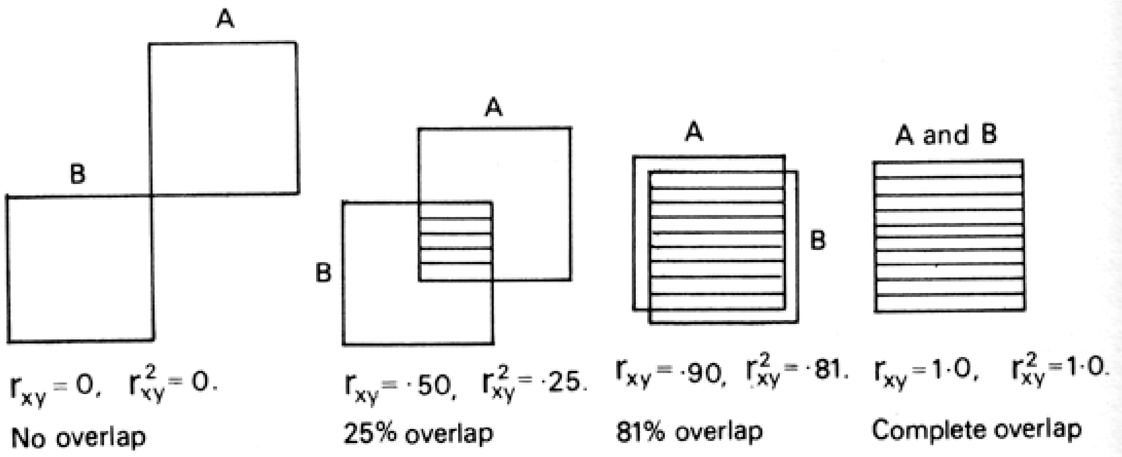

A
B
$r_{xy} = 0,\ r_{xy}^2 = 0.$
No overlap
A
B
$r_{xy} = \cdot 50,\ r_{xy}^2 = \cdot 25.$
25% overlap
A
B
$r_{xy} = \cdot 90,\ r_{xy}^2 = \cdot 81.$
81% overlap
A and B
$r_{xy} = 1\cdot 0,\ r_{xy}^2 = 1\cdot 0.$
Complete overlap

Figure 6.1 Overlap between variances of two correlated variables with changes in the value of r_{xy}

3. *The Coefficient of Alienation*

It has been shown that r_{xy}^{2} is equivalent to the proportion of explained or accounted for variance, so r_{xy}^{2} must be the square root of this proportion.

$$r_{xy} = \sqrt{\frac{\text{Explained Variance}}{\text{Total Variance}}} \qquad (6{:}6)$$

The correlation coefficient is an indicator of the degree of relationship between two variables. An index of the degree of lack of relationship is also available. It is the square root of the proportion of unexplained variance and is called the coefficient of alienation.

$$\text{Coefficient of alienation} = \sqrt{\frac{\text{Unexplained Variance}}{\text{Total Variance}}}$$
$$= \sqrt{1 - r_{xy}^{2}}$$

This sometimes provides a useful corrective to over-enthusiasm about a given value of r_{xy}. Table 6:3 shows the relationships between values of r_{xy}. and the coefficients of determination and alienation. It will be seen that not until r_{xy}. is over.70 is the degree of relationship larger than the degree of lack of relationships, and that it takes a correlation coefficient of over .70 to account for 50 per cent of the variance.

TABLE 6:3 COEFFICIENTS OF DETERMINATION AND ALIENATION AS A FUNCTION OF r_{xy}.

r_{xy}.	*Coefficient of Determination*	*Coefficient of Alienation*
.00	.00	1.00
.10	.01	.99
.20	.04	.98
.30	.09	.95
.40	.16	.92
.50	.25	.87
.60	.36	.80
.70	.49	.71
.80	.64	.60
.90	.81	.44
1.00	1.00	.00

It is apparent from this table that in terms of variance accounted for a correlation of .40 is not twice as large as one of .20. A correlation of .20 accounts for 4 per cent of the variance while a correlation of .40 accounts for four times as much.

Problems

A. What value of r_{xy}. accounts for nine times as much variance as an r_{xy}. pf /30?

B. In terms of variance accounted for, what correlation is one hundredth of the size of a correlation of 1.0?

C. What is the value of r_{xy}. when 75 per cent of the variance is not accounted for?

D. What is the value of the coefficient of alienation when 9 per cent of the variance is accounted for?

Answers

A. .90; B. .10; C. .50; D. .95.

Partial and Part correlation

1. Partial correlation

Sometimes it is desirable to know the relationship between two variables with the effects of a third variable held constant. As an example suppose that it has been demonstrated that both intelligence and number of hours worked are correlated with exam marks, and further that intelligence and number of hours worked are also correlated. All of these correlations are positive. The more intelligent tend to obtain higher exam scores and tend to work harder, those who work harder tend to be more intelligent and obtain higher exam marks. In a situation like this a straight forward correlation between intelligence and exam marks will also reflect the effect of hours worked on both intelligence and exam marks. Clearly it would be useful for us to be able to find the 'pure' correlation between intelligence and exam marks with hours worked held constant. 'Holding constant' in this situation is known as partialling out, and the technique for partialling out the effects of one or more variables from two others, in order to find the relationship between them is called partial correlation.

For this chapter and the next one a change in subscripts is desirable. Instead of using letters as subscripts with correlation coefficients it will be more useful to refer to the variables being correlated as 1, 2, 3, etc. r_{12} will be the correlation between variables 1 and 2, r_{14} between the first and fourth variable and r_{1n} between the first and *n*th variable.

Suppose that we have three variables 1, 2, and 3 and we wish to find the relationship between 1 and 2, with the effects of 3 partialled out from both. In fact what we want to do is correlate the residual scores of 1 and 2, after the parts of 1 and 2 predictable from 3 have been subtracted. It has been shown previously that the predicted Z_1 will be $r_{13}Z_3$. The symbol for Z_1 with the effects of 3 partialled out will be $Z_{1.3}$, and generally Z on variable X with Y partialled out will be $Z_{X.Y.}$ The residual Z score on 2 will be $Z_2 - r_{23}Z_3 = Z_{2.3}$. (Note that in all of these cases the subscript of the variable partialled out comes after the dot.)

The partial correlation coefficient $r_{12.3}$ will be:

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}} \qquad (7:1)$$

Proof

It is desired to correlate the residual scores $\left(Z_1 - r_{13}Z_3\right)$ and $\left(Z_2 - r_{23}Z_3\right)$ and the formula for the correlation coefficient is

$$\frac{\sum xy}{N\sigma_x\sigma_y}$$

So it will be necessary to work out

(a) the covariance of the residual scores and

(b) the standard deviation of the residual scores.

The covariance will be worked out first.

(1) $$\sum\left(Z_1 - r_{13}Z_3\right)\left(Z_2 - r_{23}Z_3\right)/N = \sum\left(Z_1Z_2 + r_{13}r_{23}Z_3{}^2 - r_{23}Z_1Z_3 - r_{13}Z_2Z_3\right)/N$$

(2) $$= \frac{\sum Z_1Z_2}{N} + r_{13}r_{23}\frac{\sum Z^2{}_3}{N} - r_{23}\frac{\sum Z_1Z_3}{N} - r_{13}\frac{\sum Z_2Z_3}{N}$$

(3) Recalling that $\frac{\sum Z_1 Z_2}{N} = r_{12} etc.$, and that $\frac{\sum Z_3{}^2}{N} = \sigma_Z^2 =$ 1.0 etc. (2) becomes: $r_{12} + r_{13} r_{23} - r_{23} r_{13} - r_{13} r_{23}$

(4) this equals $r_{12} - r_{13} r_{23}$

Turning now to the standard deviation of the residual scores we have

(5) $$\sigma_{Z1.3} = \sqrt{\frac{\sum \left(Z_1 - r_{13} Z_3\right)^2}{N}}$$

(6) $$= \sqrt{\frac{\sum Z_1{}^2}{N} + r_{13}^2 \frac{\sum Z_3^2}{N} - 2r_{13} \frac{\sum Z_1 Z_3}{N}}$$

(7) $$\frac{\sum Z_1^2}{N} = 1.0; \frac{\sum Z_2^3}{N} = 1.0; \textit{and } \frac{\sum Z_1 Z_3}{N} = r_{13}$$

So (6) becomes $\sqrt{1 + r_{13}^2 - 2r_{13}^2} = \sqrt{1 - r_{13}^2}$

(8) Repeating these steps for $Z_{2.3} = \sqrt{\frac{\sum \left(Z_2 - r_{23} Z_3\right)^2}{N}}$ gives

$\sqrt{1 - r_{23}^2}$

(9) Putting (4) as the numerator and the products of (7) and (8) as the denominator gives:-

$$r_{12.3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}$$

The other formulae for the 3 variable case are:

(a) $$r_{13.2} = \frac{r_{13} - r_{12} r_{23}}{\sqrt{1 - r_{12}^2}\sqrt{1 - r_{23}^2}} \qquad (7:2)$$

(b) $$r_{23.1} = \frac{r_{23} - r_{12} r_{13}}{\sqrt{1 - r_{12}^2}\sqrt{1 - r_{13}^2}}$$

In effect the partial correlation coefficient $r_{12.3}$ tells us what the relationship between variables 1 and 2 would be if everyone obtained the same score on variable 3.

Problems

A. Calling exam marks (1), intelligence (2) and hours worked (3), and given r_{12} = .50, and r_{13}=.40, and r_{23} of .40 work out the value of $r_{12.3}$.

B. Given three variables (1) prognosis in terms of weeks to recover, (2) an anxiety questionnaire, (3) a physiological measure, and $r_{12} = .40; r_{13} = .30,\ and\ r_{23} = .80$, what is the correlation of the physiological measure with prognosis with the anxiety questionnaire results partialled out from both variables?

Answers

A. $r_{12.3} = \dfrac{.50-(.40\times.40)}{\sqrt{1-.40^2}\sqrt{1-.40^2}} = \dfrac{.34}{.86} = .396$

B. $r_{13.2} = \mathit{from}\ (7:2)\dfrac{r_{13}-r_{12}r_{23}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{23}^2}}$

$$= \frac{.30-(.40\times.80)}{\sqrt{1-.40^2}\sqrt{1-.80^2}} = \frac{-.02}{.93\times.60} = -.036$$

A partial correlation coefficient with one variable partialled out is called a first order partial *r*, with two variables partialled out a second order partial *r* and so on. The general formula for a second order *r* is:

$$r_{12.34} = \frac{r_{12.3}-r_{14.3}r_{24.3}}{\sqrt{1-r_{14.3}^2}\sqrt{1-r_{24.3}^2}} \qquad (7:3)$$

For an *N*th order partial the formula is:

$$r_{12.34...N} = \frac{r_{12.34...(N-1)}-r_{1N.34...(N-1)}r_{2N.34...(N-1)}}{\sqrt{1-r_{1N.34...(N-1)}^2}\sqrt{1-r_{2N.34...(N-1)}^2}} \qquad (7:4)$$

Partial correlation assumes linearity of regression between all variables. If there are serious departures from linearity a partial *r* will be meaningless.

2. *Part or Semi-partial Correlation*

In the case of partial correlation the variable partialled out is partialled out from both of the variables of interest. However, it is also possible to correlate partialled scores on one variable with ordinary scores on another. This type of correlation is called part or semi-partial correlation. The formula for the part correlation coefficient is:

Part correlation coefficient: $r_{1(2.3)} = \dfrac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{23}^2}}$ (7:5)

Note that the partialled variable and the variable partialled from it are put in brackets so $r_{1(2.3)}$ is the correlation between 1 and 2 with the effects of 3 partialled out from 2.

Proof

(1) $$r_{1(2,3)} = \frac{\sum Z_1(Z_2 - r_{23}Z_3)}{N\sqrt{1-r_{23}^2}}$$

(The standard deviation of the Z_1 scores will be 1 and it has been shown in 7:1 step 8 that the standard deviation of

$$[Z_2 - r_{23}Z_3] = \sqrt{1-r_{23}^2} \text{)}$$

(2) $$r_{1(2,3)} = \frac{\sum Z_1Z_2 - r_{23}\sum Z_1Z_3}{N\sqrt{1-r_{23}^2}}$$

(3) Dividing numerator and denominator by *N* gives:

$$r_{1(2,3)} = \frac{r_{12} - r_{23}r_{13}}{\sqrt{1 - r_{23}^2}}$$

$$\text{(because } \frac{\sum Z_1 Z_2}{N} = r_{12} \; and \; \frac{\sum Z_1 Z_3}{N} = r_{13})$$

Other formulae for the three variable case include:

(a) $$r_{2(1.3)} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2}} \qquad (7{:}6)$$

(b) $$r_{3(1.2)} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{1 - r_{13}^2}}$$

Part or semi-partial correlation has the effect of reducing the correlation between the partialled variable and the variable partialled from it to zero.

$$r_{3(1.3)} = r_{1(2.1)} = r_{2(1.2)} \; etc. = 0 \qquad (7{:}7)$$

Proof

(1) $$r_{1(2.1)} = r_{12} - r_{12}r_{11}$$

(2) $$r_{11} = \frac{\sum Z_1 Z_1}{N} = \frac{\sum Z_1^{\;2}}{N} = 1.0$$

(3) So (1) becomes

$$r_{1(2.1)} = \frac{r_{12} - r_{12}}{\sqrt{1 - r_{12}^2}} = 0$$

Both part and partial correlation are useful in prediction problems, and have a fairly straightforward relationship to multiple correlations, as will be seen in the next chapter.

Problems

A. Given the following data on the relationship between prognosis (1), anxiety questionnaire (2), and physiological measure (3), r_{12} = .40 r_{13} = .30, and r_{23} = .80. What is the correlation between the physiological measure and prognosis with anxiety questionnaire scores partialled out from the physiological measure i.e. what is the value of $r_{1(3.2)}$?

B. A performance test (1) and a verbal intelligence test (2) are used for predicting scholastic success (3). You want to know the correlation between the performance test, with verbal intelligence partialled out from it, and exam marks. If r_{12} =.60, r_{13} = .60, and r_{23} = .40, what will the correlation be?

Answers

A.
$$r_{1(3.2)} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{1-r_{23}^2}} = \frac{.30 - (.40 \times .80)}{\sqrt{1-.80^2}} = \frac{-.02}{.60}$$
$$= -.033$$

B.
$$r_{3(1.2)} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{1-r_{13}^2}} = \frac{.60 - (.60 \times .40)}{\sqrt{1-.60^2}} = \frac{.36}{.80}$$
$$= +.45$$

3. *The Partial Standard Deviation*

As has been stated previously, scores from which another variable has been partialled out are called partial scores or residual scores. Partial scores have been used in our discussion of the partial and part correlation. The formula for a partial score with one variable partialled out is:

$$\text{Partial score} = Z_{1.2} = Z_1 - r_{12}Z_2 \qquad (7:8)$$

The standard deviation of these partial scores has also been used and derived in 7.1, (5 to 8).

$$\text{Partial standard deviation} = \sigma_{Z1.2} = \sqrt{1-r_{12}^2} \qquad (7:9)$$

In terms of raw scores (7:8) becomes:

Partial score: $$Y.X = Y - r_{xy}\frac{\sigma_y}{\sigma_x}\left(X - M_x\right) + M_y \qquad (7:10)$$

or in terms of deviation scores:

Partial score: $$y.x = y - r_{xy}\frac{\sigma_y}{\sigma_x}x \qquad (7:11)$$

The raw score standard deviation is.

$$\sigma_{y.x} = \sigma_y\sqrt{1 - r_{xy}^2} \qquad (7:12)$$

Recalling that r_{xy}^2 is the coefficient of determination, and that the coefficient of determination is the proportion of variance accounted for, it can be seen from (7:9) that the partial standard deviation is the square root of the variance remaining after variance attributable to another variable has been subtracted. To convert this to a raw score form as in (7:12) $\sqrt{1-r_{xy}^2}$ is multiplied by the standard deviation of the partialled variable.

Higher order partial standard deviations are also equal to the square root of the variance remaining after the effects of the other variables have been partialled out. Let us consider the case of $\sigma_{1.23}$, which is the partial standard deviation of variable 1 with variables 2 and 3 partialled out. The proportion of variance in variable 1 accounted for by variable 2 will be r_{12}^2. Thus after variable 2 has been partialled out the proportion of variance remaining will equal $1-r_{12}^2$. Of this remainder some will be accounted for by variable 3. If variable 2 and variable 3 were not related the variance attributable to variable 3 would be r_{13}^2. The variance remaining after the partialling out of variables 2 and 3 would, therefore, be $1-r_{12}^2-r_{13}^2$. The square root of this would be the partial standard deviation.

The situation becomes more complicated when variables 2 and 3 correlated with one another. Starting as before, the proportion of variance accounted for by variable 2 will be r_{12}^2, and as before the remaining variance will equal $1-r_{12}^2$.

However, because 2 and 3 are correlated it will not be possible merely to subtract r_{13}^2 as the proportion of variance attributable to variable 3. As variables 2 and 3 are correlated some of the variance in 1 accounted for by 2, will be shared with variable 3. To obtain the proportion of variance in 1 accounted for by 3, from which variance also accounted for by 2 is excluded, it is necessary to use the partial correlation coefficient $r_{13.2}$. The square of this will give the proportion of variance in 1 which is attributable to 3 after the effects of 2 have been excluded from 1 and 3.

Therefore:

(a) the variance remaining after variable 2 has been partialled out will be:-

$$1-r_{12}^2$$

(b) of this remainder, variable 3 will account for a proportion of $r_{13.2}^2$, leaving a remainder of

$$1-r_{13.2}^2$$

Multiplying (a) and (b) together will give the variance remaining after both 2 and 3 have been partialled out. So the formula for the variance not accounted for by the partialling out of the two variables will be:

Partial variance with two variables partialled out

$$(Z\text{ score form}) = \sigma^2_{Z1.23} = \left(1-r^2_{12}\right)\left(1-r^2_{13.2}\right) \tag{7:13}$$

$$\sigma_{Z1.23} = \sqrt{\left(1-r^2_{12}\right)\left(1-r^2_{13.2}\right)} \tag{7:14}$$

Raw score form

$$\sigma_{1.23} = \sigma_1\sqrt{\left(1-r^2_{12}\right)}\left(1-r^2_{13.2}\right) \tag{7:15}$$

Suppose that $r_{12} = .71$, and $r_{13.2} = .50$. The proportion of Variance accounted for by $r_{12} = r^2_{12} = .50$. Fifty per cent of the variance is accounted for which leaves 50 per cent not accounted for. Of this remaining 50 per cent, 25 per cent can be accounted for by variable 3 with 2 partialled out from it. Thus 75 per cent of the 50 per cent remaining after the effects of variable 2 have been allowed for will still remain unaccounted for after variable 3 has been partialled out.

The partial standard deviation will therefore be $\sqrt{.50 \times .75} = \sqrt{.375.}$ This is the value which would also be obtained by use of formula (7:13).

By similar reasoning it can be shown that the partial standard deviation of variable 1 with variables 234...*N* partialled out is:

$$\sigma_{Z1.234\ldots N} \tag{7:16}$$

$$= \sqrt{\left(1-r^2_{12}\right)\left(1-r^2_{13.2}\right)\left(1-r^2_{14.23}\right)\cdots\left[1-r^2_{1N.234\ldots(N-1)}\right]}$$

The raw score equivalent can be obtained by multiplying by σ_1.

Problem

Given $r_{12} = .40, r_{13} = .50,$ and $r_{23} = .60,$ what is the value of $\sigma_{Z1.23}$?

Answer

(a) Two correlation coefficients are needed r_{12} and $r_{13.2}$. The formula for the latter will be:

$$\frac{r_{13} - r_{12}r_{23}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{23}^2}} = \frac{.50-(.40\times.60)}{\sqrt{1-.16}\sqrt{1-.36}} = \frac{.26}{.92\times.80} = .35$$

(b) The formula for the partial standard deviation thus becomes:

$$\sqrt{(1-.40^2)(1-.35^2)} = \sqrt{.74} = .86$$

Multiple regression and prediction

1. The Multiple Regression Equation

In many applied situations there are a number of variables correlated with a criterion and the problem arises of how best to weight them to obtain the most accurate prediction of the criterion. If it can be assumed that the regressions of the variables on one another are linear, then the usual technique is to use a multiple linear regression equation. It will be recalled that in the two variable case the linear regression equation was:

$$\hat{Y} = a + bX$$

In the multi-variate case the multiple regression equation is:

$$\hat{Y} = a + b_1 X_1 + b_2 X_2 + \ldots b_n X_n \qquad (8:1)$$

Just as the weights in the two variables case were chosen to minimize the sum of squared deviations between predicted and obtained scores - the principle of least squares – so in the multi-variate case. Weights are chosen to make the value of:

$$\sum \left(Y - a - b_1 X_1 - b_2 X_2 - \ldots b_n X_n\right)^2$$

as small as possible.

The weights assigned to the various predictor variables will be determined by:

(a) the correlation between the predictor variable and the criterion; and

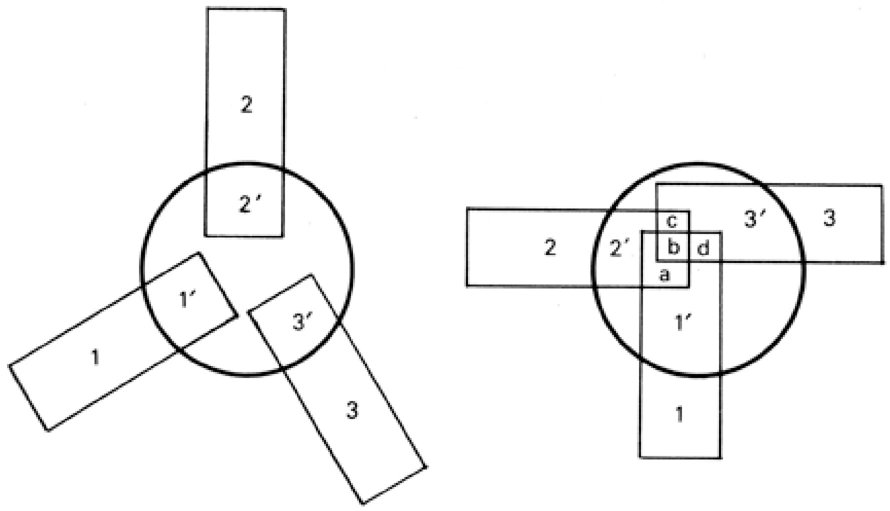
(b) the intercorrelations of the predictor variables.

The first of these is probably fairly obvious. In general the higher the correlation between a predictor variable and the criterion the higher its weight would be expected to be. The second is also fairly clear. Ideally we would like predictor variables with:

(a) high correlations with the criterion; and

(b) low correlations with each other.

This is because the higher the correlations between the predictor variables, the more variance they share in common and the more likely it becomes that parts of the criterion variance accounted for by each will overlap with criterion variance accounted for by the other. These considerations are presented diagrammatically in Figure 8.1, which shows two situations. In 8.1(a) there is no correlation between predictors and in 8.1(b) the predictors are correlated. In the diagram the predictor variables are shown as oblongs 1, 2, and 3, and the criterion as a circle, O.



(a) Zero correlation between predictors (b) Non-zero correlation between predictors

Figure 8.1 Relationships between predictors (1, 2, 3) and a criterion (0)

Turning to 8.1(a) first it will be seen that the amount of the criterion covered by the three predictors is 1′ + 2′ + 3′, which are the parts of the predictors overlapping the criterion. In 8.1(b) however, the sum of the overlaps (1′ + 2′ + 3′) would overestimate the total overlap because of the overlap between the predictors themselves. In 8.1(b) the amount of the criterion covered will be:

1′ + (2′ – a – b) + (3′ – b – c - d)

So to obtain total overlap it has been necessary to add 1 + 2 minus the overlap of 2 and 1, and 3 minus the overlap of 3 with 1 and with 2.

In effect we have partialled out the effects of 2 from 1; and of 1 and 2 from 3. If in the above argument we replace ‘overlap’ by ‘proportion of variance accounted for’ then in 8.1(a)

(1) overlap 1′ will equal r_{01}^2

(2) overlap 2′ will equal r_{02}^2

(3) overlap 3′ will equal r_{03}^2

Total overlap will equal the sum of these, i.e. $r_{01}^2 + r_{02}^2 + r_{03}^2$, and will be equal to the proportion of variance accounted for by the three predictors.

In the two variable case it will be recalled that the proportion of variance accounted for equalled the coefficient of determination-r_{xy}^2-and the square root of this coefficient of determination was the correlation coefficient r_{xy}. In the present situation we have a proportion of variance accounted for by a number of predictors, which is the coefficient of multiple determination – symbolised in this case as $R_{0.123}^2$. Note the use of a capital *R* as opposed to a small *r* in the bivariate case, and the subscripts. In the subscripts the criterion variable appears before the dot and the predictor variables after it. As in the bivariate case the square root of the coefficient of determination is a correlation coefficient, so in 8.1(a) or any case where the predictor variables are uncorrelated the multiple correlation coefficient will be:

$$R_{0.12...n} = \sqrt{r_{01}^2 + r_{02}^2 + ...r_{0n}^2} \qquad (8:2)$$

It must be re-emphasized that this is the multiple correlation coefficient when the predictor variables are uncorrelated.

In 8.1(b) the predictor variables are correlated. It was, therefore, necessary to partial out the overlap and we obtained $1' + (2' - a - b) + (3' - b - c - d)$, where *a* + *b* was the overlap of 1 and 2; and *b* + *c* + *d* was the overlap between 1, 2 and 3. This is, therefore, a part correlation problem, and the proportion of variance accounted for will be given by:

$$R_{0.12...n} = \sqrt{r_{01}^2 + r_{0(2.1)}^2 + ...r_{0(n.12...[n-1])}^2} \qquad (8:4)$$

Other formulae are available, one of which will be examined later, after the weights used in the multiple regression equation have been considered.

2. *The Weights in the Multiple Regression Equation*

At this stage it will be useful to recall that in the two variable case the weight used with *X* for prediciting *Y* from *X* was $r_{xy}\frac{\sigma_y}{\sigma_x}$; and for predicting *X* from *Y* was $r_{xy}\frac{\sigma_x}{\sigma_y}$. These can be seen to be the correlation coefficient multiplied by the ratio of the standard deviation of the criterion to the standard deviation of the predictor. In the multiple regression equation the weights will be analogous but because a partialling out has to be done the correlation coefficient will be a partial one and the standard deviations will be partial ones. As the derivation of the weights involves more than simple algebra it will not be described, but it can be shown that:

(a) $$b_{01.2} = r_{01.2}\frac{\sigma_{0.2}}{\sigma_{1.2}} \qquad (8.5)$$

(b) $$b_{01.23} = r_{01.23}\frac{\sigma_{0.23}}{\sigma_{1.23}}$$

(c) $$b_{01.23...n} = r_{01.23...n}\frac{\sigma_{0.23...n}}{\sigma_{1.23...n}}$$

It is common practice to economise on subscripts by writing b_1 for the weight to be attached to variable l, b_2 for that attached to variable 2 and so on. If the scores on the predictors are in *Z* score form then β's (betas), are used instead of *b*'s. β's and *b*'s will not have the same value but there is a simple relationship between them:

(a) $$\beta_{01.2} = b_{01.2}\frac{\sigma_1}{\sigma_0} \qquad (8:6)$$

(b) $$b_{01.2} = \beta_{01.2} \frac{\sigma_0}{\sigma_1}$$

(c) $$b_{01.2} = \beta_{01.2} \frac{\sigma_0}{\sigma_1}$$

(d) $$b_{01.23 \ldots n} = \beta_{01.23 \ldots n} \frac{\sigma_0}{\sigma_1}$$

As in the case of *b*'s, $\beta_{01.23 \ldots n}$ is often written simply as β_1, and similarly other β's are commonly written as β_2, β_n, etc.

Recalling the formulae for partial correlation coefficients and partial standard deviations as described in the last chapter, it will be apparent that the computational labour involved in working out multiple regression weights is considerable, and best left to a computer. However, to give more concrete practice with multiple regression the next section will consider some aspects of multiple prediction in the case where there are only two predictors.

3. *Prediction with Two Predictors*

Let us start by examining the prediction of $\hat{Z}_0$ from Z_1 and Z_2. The formula will be:

$$\hat{Z}_0 = \beta_1 Z_1 + \beta_2 Z_2 \qquad (8:7)$$

From (8:6) we know that $\beta_1 = b_1 \dfrac{\sigma_1}{\sigma_0}$. This equals

$$r_{01.2}\frac{\sigma_0\sqrt{1-r_{02}^2}}{\sigma_1\sqrt{1-r_{12}^2}}.\frac{\sigma_1}{\sigma_0} = r_{01.2}\frac{\sqrt{1-r_{02}^2}}{\sqrt{1-r_{12}^2}} \text{ therefore:}$$

$$\beta_1 = \frac{r_{01}-r_{02}r_{12}}{\sqrt{1-r_{02}^2}\sqrt{1-r_{12}^2}}.\frac{\sqrt{1-r_{02}^2}}{\sqrt{1-r_{12}^2}} \qquad (8:8)$$

Which simplifies to:

$$\frac{r_{01}-r_{02}r_{12}}{1-r_{12}^2}$$

Similarly:

$$\beta_2 = \frac{r_{02}-r_{01}r_{12}}{\sqrt{1-r_{01}^2}}.\frac{\sqrt{1-r_{01}^2}}{\sqrt{1-r_{12}^2}} = \frac{r_{02}-r_{01}r_{12}}{1-r_{12}^2} \qquad (8:9)$$

Note that when r_{12} equals zero, β_1 becomes r_{01}; and β_2 becomes r_{02}. This makes (8:7) equal:

$$Z_0 = r_{01}Z_1 + r_{02}Z_2$$

Problems

A. Given the following data:

$$r_{01} = .40;\ r_{02} = .60;\ r_{12} = .71;\ Z_1 = +1.0;\ Z_2 = -2.0;$$

What are the values of:

(1) β_1?

(2) β_2?

(3) $\hat{Z}_0$?

B. If r_{12} had been equal to zero in problem (a), what would have been the values of:

(1) β_1?

(2) β_2?

(3) $\hat{Z}_0$?

Answers

A. (1) -.05; (2) +.63; (3) -1.31

B. (1) .40; (2) .60; (3) -0.8

The derivation of the raw score multiple regression equation from the standard $\hat{X}_0 - M_0 = b_1(X_1 - M_1) + b_2(X_2 - M_2)$ score version is fairly simple.

$$\hat{X}_0 = M_0 - b_1M_1 - b_2M_2 + b_1X_1 + b_2X_2 \tag{8:10}$$

Proof

(1) Given that $Z_0 = \beta_1Z_1 + \beta_2Z_2$

(2) Changing the Z scores to raw score equivalents gives:

$$\frac{\hat{X}_0 - M_0}{\sigma_0} = \beta_1\left(\frac{X_1 - M_1}{\sigma_1}\right) + \beta_2\left(\frac{X_2 - M_2}{\sigma_2}\right)$$

(3) Multiplying both sides by σ_0 gives:

$$\hat{X}_0 - M_0 = \beta_1\frac{\sigma_0}{\sigma_1}(X_1 - M_1) + \beta_2\frac{\sigma_0}{\sigma_2}(X_2 - M_2)$$

(4) But $\beta_1\frac{\sigma_0}{\sigma_1}$ and $\beta_2\frac{\sigma_0}{\sigma_2}$ have been defined as b_1 and b_2 respectively (see (8:6)) so (3) becomes:

(5) Adding M_0 to both sides and rearranging gives:

$$\hat{X}_0 = M_0 - b_1M_1 - b_2M_2 + b_1X_1 + b_2X_2$$

Problem

Given the data in section (a) of the previous example and that $M_0 = 100; M_1 = 100;$ and $M_2 = 50;$ and that $\sigma_0 = 10;\ \sigma_1 = 20;\ \sigma_2 = 10$, what will be the values of:

(1) b_1?

(2) b_2?

(3) $\hat{X}_0$?

Answers

(1) $b_1 = \beta_1 \dfrac{\sigma_0}{\sigma_1} = -.05\left(\dfrac{10}{20}\right) = -.025$

(2) $b_2 = \beta_2 \dfrac{\sigma_0}{\sigma_2} = .63\left(\dfrac{10}{10}\right) = .63$

(3) $M_0 - b_1M_1 - b_2M_2 + b_1X_1 + b_2X_2$

$$= 100 - (-.025 \times 100) - (.63 \times 50) + (-.025 \times 120)$$
$$+ (.63 \times 30) = 86.9$$

4. *The Multiple Correlation Coefficient in Terms of Beta Weights*

An alternative formula for the multiple correlation coefficient is:

$$R_{0.12\ldots n} = \sqrt{\beta_1 r_{01} + \beta_2 r_{02} + \ldots \beta_n r_{0n}} \qquad (8:11)$$

This formula looks very different from the one earlier in the chapter which defined the multiple correlation coefficient in terms of part correlation, (Formula (8:3)), but in fact the two are the same. This will be shown to be true for the case of two predictor variables.

$$\sqrt{\beta_1 r_{01} + \beta_2 r_{02}} = \sqrt{r_{01}^2 + r_{0(2.1)}^2} \qquad (8:12)$$

Proof

(1) $\beta_1 = \frac{r_{o1} - r_{02}r_{12}}{1 - r_{12}^2}$ *and* $\beta_2 = \frac{r_{02} - r_{01}r_{12}}{1 - r_{12}^2}$

(2) Therefore $\beta_1 r_{01} = \frac{r_{01}^2 - r_{01}r_{02}r_{12}}{1 - r_{12}^2}$ *and*

$$\beta_2 r_{02} = \frac{r_{02}^2 - r_{02}r_{01}r_{12}}{1 - r_{12}^2}$$

(3) Therefore $\beta_1 r_{01} + \beta_2 r_{02} = \frac{r_{01}^2 + r_{02}^2 - 2r_{01}r_{02}r_{12}}{1 - r_{12}^2}$

(4) By adding *and* subtracting $r_{01}^2 r_{12}^2$ to the numerator we obtain:

$$\beta_1 r_{01} + \beta_2 r_{02} = \frac{r_{01}^2 + r_{02}^2 - 2r_{01}r_{02}r_{12} + r_{01}^2 r_{12}^2 - r_{01}^2 r_{12}^2}{1 - r_{12}^2}$$

(5) (4) can be split into two parts:

$$\frac{r_{01}^2 - r_{01}^2 r_{12}^2}{1 - r_{12}^2} + \frac{r_{02}^2 + r_{01}^2 r_{12}^2 - 2r_{01}r_{02}r_{12}}{1 - r_{12}^2}$$

(6) $$\frac{r_{01}^2 - r_{01}^2 r_{12}^2}{1 - r_{12}^2} = \frac{r_{01}^2 \left(1 - r_{12}^2\right)}{1 - r_{12}^2} = r_{01}^2$$

(7) So (5) equals:

$$r_{01}^2 + \frac{r_{02}^2 + r_{01}^2 r_{12}^2 - 2r_{01}r_{02}r_{12}}{1 - r_{12}^2}$$

(8) However:

$$\frac{r_{02}^2 + r_{01}^2 r_{12}^2 - 2r_{01}r_{02}r_{12}}{1 - r_{12}^2} = \left(\frac{r_{02} - r_{01}r_{12}}{\sqrt{1 - r_{12}^2}}\right)^2$$

(9) From Formula (7:6):

$$\left(\frac{r_{02} - r_{01}r_{12}}{\sqrt{1 - r_{12}^2}}\right)^2 = r_{0(2.1)}^2$$

(10) Substituting this value in (7) gives:

$$\sqrt{\beta_1 r_{01} + \beta_2 r_{02}} = \sqrt{r_{01}^2 + r_{0(2.1)}^2}$$

5. *The Standard Deviation of the Predicted Scores*

The variable $\hat{X}_0$ obtained by using a multiple regression equation will have a smaller standard deviation than the actual criterion variable. The formula relating these is:

$$\hat{\sigma}_0 = R_{0.12...n}\sigma_0 \qquad (8:13)$$

Where:

$\hat{\sigma}_0 =$ standard deviation of the predicted scores, and
$\sigma_0 =$ standard deviation of the criterion scores

Proof

(1) $R_{0.12...n}$ is the correlation coefficient between predicted and obtained scores, therefore:

$$\hat{x}_0 = R_{0.12...n}\frac{\hat{\sigma}_0}{\sigma_0}x_0$$

(2) Translating to *Z* scores by dividing both sides by $\hat{\sigma}_0$ and symbolizing $R_{0.12...n}$ as *R*, this becomes:

$$\hat{Z}_0 = RZ_0$$

(3) Therefore:

$$\frac{\sum\left(\hat{Z}_0 - \bar{Z}_0\right)^2}{N} = \frac{\sum\left(RZ_0 - R\bar{Z}_0\right)^2}{N}$$

(4) As mean *Z* scores equal zero (3) becomes:

$$\frac{\sum\hat{Z}_0^2}{N} = R^2\left(\frac{\sum Z_0^2}{N}\right)$$

(5) The term on the left is the variance of the $\hat{Z}_0$ Scores, and on the right is the variance of Z_0 scores multiplied by R^2. Converting to raw scores gives:

$$\hat{\sigma}_0^2 = R^2\sigma_0^2$$

(6) Therefore

$$\hat{\sigma}_0 = R\sigma_0 = R_{0.12...n}\sigma_0$$

6. *The Interpretation of a Multiple Correlation Coefficient*

The multiple correlation coefficient is interpreted in a similar way to the ordinary correlation coefficient. The coefficient of multiple determination gives the proportion of variance accounted for.

$R^2_{0.12\ldots n}$ = coefficient of multiple determination. (8:14)

Also analogous to the bivariate case is the standard error of multiple estimate.

$\sigma_0\sqrt{1-R^2_{0.12\ldots n}}$ = standard error of multiple estimate. (8:15)

This gives the standard deviation of scores around the regression line.

It is also possible to use a coefficient of multiple alienation, or multiple non determination.

Problems

A. If Test 1 is a test of extraversion and Test 2 WAIS IQ, and the correlation between intelligence and extraversion is zero; between extraversion and exam marks is -.50; and between IQ and exam marks is + .60:

What are the values of

(1) β_1?

(2) β_2?

(3) $R_{0.12}$?

B. If student A obtains a score at the 84th percentile for extraversion. And a WAIS IQ of 130, what is his predicted *Z* score for exam marks?

C. Would he do better in examinations than student B whose score lies at the 50th percentile for extraversion, and whose IQ is 120?

D. If extraversion and intelligence were correlated -.40, what would be the values of:

(1) β_1?

(2) β_2?

(3) $R_{0.12}$?

Answers

A. In problem (A)r_{12} = -.50

1 = r_{01} = -.50

2 = r_{02} = +.60

and $r_{0.12} = \sqrt{r_{01}^2 + r_{02}^2} = \sqrt{(-50)^2 + (.60)^2} = .78$

B. $Z_1 = 1;\ Z_2 = 2;\ Z_0 = \beta_1 Z_1 + \beta_2 Z_2 = (-.50 \times 1) + (.60 \times 2) = +.70$

C. From Answer (B) we know that student A obtains a $\hat{Z}_0$ score of +.70. Student B will obtain a $\hat{Z}_0$ score of
$(-.50 \times 0) + (.60 \times 1.33) = +.80$.
The answer is therefore 'No'.

D. (1) $\beta_1 = \dfrac{-.50-(-.40\times .60)}{1-(-.40)^2} = \dfrac{-.26}{.84} = -.31$

(2) $\beta_2 = \dfrac{.60-(-.50\times -.40)}{1-(-.40)^2} \quad \dfrac{.40}{.84} = +.48$

(3) $R_{0.12} = \sqrt{\beta_1 r_{01} + \beta_2 r_{02}}$

$= \sqrt{(-.31\times -.50)} + (.48\times .60)$

$\sqrt{.44} = .66$

Composite scores

1. *The Mean of Composite Scores*

A composite score is the score which results from summing two or more scores. Composite scores will be symbolised as C.

It can be shown that the mean of a composite equals the sum of the means of the components. Using $\overline{C}$ as the mean of the composite and

$\overline{X}_1, \overline{X}_{2,}$ etc. for the means of the components we have:

$$\overline{C} = \overline{X}_1 + \overline{X}_2 + \ldots \overline{X}_n \qquad (9:1)$$

Proof

(1) $\overline{C} = \frac{\sum C}{N}$

(2) $\sum C = \sum (X_2 + X_2 + \ldots X_n)$

(3) By Summation Rule 1. (2) is equal to

$$\sum X_1 + \sum X_2 + \ldots \sum X_n.$$

(4) Therefore:

$$\frac{\sum C}{N} = \frac{\sum X_1}{N} + \frac{\sum X_2}{N} + \ldots \frac{\sum X_n}{N}$$

(5) The values on the right are of course means so

$$\overline{C} = \overline{X}_1 + \overline{X}_2 + ...\overline{X}_n$$

S_{xy}

Sometimes a composite score is the sum of weighted components. For example *C* might equal $2X_1 + 1.5X_2 + kX_3$, the weights being 2 for *X,* 1.5 for X_2, and *k* for X_3. Let us symbolise weights as W_1, $W_2 \ldots W_n$, then:

$$\overline{C} = W_1\overline{X}_1 + W_2\overline{X}_2 + ...W_n\overline{X}_n \qquad (9:2)$$

Proof

(1) $C = (W_1X_1 + W_2X_2 + ...W_nX_n)$

(2) $\sum C =$ (using Summation Rule 1)

$$\sum W_1X_1 + W_2X_2 + ...\sum W_nX_n$$

(3) So $\dfrac{\sum C}{N} = \dfrac{\sum W_1X_1}{N} + \dfrac{\sum W_2X_2}{N} + ...\dfrac{\sum W_nX_n}{N}$

(4) So $\overline{C} = W_1\overline{X}_1 + W_2\overline{X}_2 + ...W_n\overline{X}_n$

2. *The Covariance*

Before considering the variance of a composite it will be worth recalling the covariance and some computational formulae connected with it. The covariance has been mentioned before in the chapter on correlation. It is defined as the mean of the products of subjects′ deviation scores on two tests. Using S_{xy} as the symbol for the covariance:

$$S_{xy} = \frac{\sum(X - M_2)(Y - M_y)}{N} = \frac{\sum xy}{N} \qquad (9{:}3)$$

It will be recalled that one formula for r_{xy} was:

$$r_{xy} = \frac{\sum xy}{N\sigma_x\sigma_y} \qquad (9{:}4)$$

From this it follows that:

$$S_{xy} = \frac{1}{N}\sum xy = r_{xy}\sigma_x\sigma_y$$

That is the mean product of the deviation scores equals the product of the correlation coefficient and the two standard deviations. A convenient formula for the covariance is:

$$S_{xy} = \frac{\sum XY}{N} - M_x M_y \qquad (9{:}6)$$

Proof

(1) $(X - M_x)(Y - M_y) = XY + M_x M_y - M_x Y - M_y X$

(2) $\sum(X - M_x)(Y - M_y) = \sum XY + NM + NM_x M_y - M_x \sum Y - M_y \sum X$

(3) Dividing by *N* gives:

$$\frac{\sum(X - M_x)(Y - M_y)}{N} = \frac{\sum XY}{N} + M_x M_y - M_x \frac{\sum Y}{N} - M_y \frac{\sum X}{N}$$

(4) But $\frac{\sum Y}{N} = M_y$; and $\frac{\sum X}{N} = M_x$

So the right hand term becomes: $\frac{\sum XY}{N} - M_x M_y$

3. *The Variance of a Composite Score*

By now the formula for the variance is familiar, i.e.

$$\frac{\sum(X - M_x)^2}{N}$$

so the formula for the variance of a composite will be:

$$\sigma_C^2 = \frac{\sum(C - \overline{C})^2}{N}$$

$$= \frac{\sum\left[(X_1 + X_2 + \ldots X_n) - (\overline{X}_1 + \overline{X}_2 + \ldots \overline{X}_n)\right]^2}{N} \qquad (9:7)$$

The last term can be written in deviation scores:

$$\sigma_C^2 = \frac{\sum\left[(X_1 + X_2 \ldots X_n) - (\overline{X}_1 + \overline{X}_2 + \ldots \overline{X}_n)\right]^2}{N} \qquad (9:8)$$

$$=\frac{\sum\left[\left(X_1-\overline{X}_1\right)+\left(X_2-\overline{X}_2\right)+\ldots\left(X_n-\overline{X}_n\right)\right]^2}{N}$$

$$=\frac{\sum\left(x_1+x_2\ldots x_n\right)^2}{N}$$

An easy way to work out all of the values involved in $\left(x_1+x_2+\ldots x_n\right)^2$ is to prepare a square table thus:

	x_1	x_2	x_3	...	x_n
x_1					
x_2					
x_3					
...					
x_n					

The body of the table is formed by multiplying the marginal elements. As follows:-

	x_1	x_2	x_3	...	x_n
x_1	x_1^2	x_1x_2	x_1x_3	...	x_1x_n
x_2	x_1x_2	x_2^2	x_2x_3	...	x_2x_n
x_3	x_1x_3	x_2x_3	x_3^2	...	x_3x_n
...	...	...	...	...	...
x_n	x_1x_n	x_2x_n	x_3x_n	...	x_n^2

Thus for each individual we have:

(a) $x_1^2+x_2^2+\ldots x_n^2$ and

(b) $2x_1x_2 + 2x_1x_3 + ...2x_{(n-1)}x_n$

Summing across individuals and dividing by *N* gives:

$$\frac{\sum(x_1 + x_2 ... x_n)^2}{N}$$

$$= \frac{\sum\left(x_1^2 + x_2^2 + ...x_n^2 + 2x_1x_2 + 2x_1x_3 + ...2x_{(n-1)}x_n\right)}{N} \qquad (9:9)$$

But by definition $\frac{\sum x_1^2}{N} = \sigma_1^2$, etc., and $\frac{\sum x_1x_2}{N} =$ covariance $x_1x_2 = S_{x_1x_2}$, etc. The variance of a composite is therefore equal to the sum of the variances of the components plus twice the sum of all possible covariances:

$$\sigma_C^2 = \sigma_{x_1}^2 + {}_{x_2}^2 + ...\sigma_{x_n}^2 + 2S_{x_1x_2} + 2\mathrm{S}_{x_1x_3} + ...2S_{x_{(n-1)}x_n} \qquad (9:10)$$

But from (9:5) $2S_{x_1x_2} = 2r_{x_1x2}\sigma_{x_1}\sigma_{x_2}$ (9:10) can be written as:

$$\sigma_C^2 = \sigma_{1_1}^2 + \sigma_{x_2}^2 + ...\sigma_{x_n}^2 + 2r_{x_1x_2}\sigma_{x_1}\sigma_{x_2} + ...2r_{x_{(n-1)}x_n}\sigma_{x_{(n-1)}}\sigma_{x_n}$$

If the variables are weighted the variance can again be worked out fairly simply by using a square table.

	W_1x_1	W_2x_2	W_3x_3	...	W_nx_n
W_1x_1	$W_1{}^2x_1{}^2$	$W_1W_2x_1x_2$	$W_1W_3x_1x_3$	...	$W_1W_nx_1x_n$
W_2x_2	$W_1W_2x_1x_2$	$W_2{}^2x_2{}^2$	$W_2W_3x_2x_3$	...	$W_2W_nx_2x_n$
W_3x_3	$W_1W_3x_1x_3$	$W_2W_3x_2x_3$	$W_3{}^2\ x_3{}^2$	...	$W_3W_nx_3x_n$
...	...	...	...	...	...
W_nx_n	$W_1W_nx_1x_n$	$W_2W_nx_2x_n$	$W_3W_nx_3x_n$	...	$W_n{}^2x_n{}^2$

The variance of a composite of weighted components will therefore be:

$$\sigma_C^2 = W_1{}^2\sigma_{x_1}^2 + W_2{}^2\sigma_{x_2}^2 + ...W_n{}^2\sigma_{x_n}^2 + 2W_1W_2S_{x_1x_2} + ...2W_{(n-1)}W_nS_{x_{(n-1)}x_n} \qquad (9:12)$$

For many purposes it is useful to look at the variances of composite variables slightly differently. Such variances are made up of a variance for each of n tests and a number of terms of the type $2r_{x_ix_j}\sigma_i\sigma_j$. The sum of the variances will be $\sum\sigma^2$. and the mean variance will be $\frac{\sum\sigma^2}{N} = \overline{\sigma}^2$. From this it follows that $\sum\sigma^2 = n\overline{\sigma}^2$. Similarly $\sum r_{ij}\sigma_i\sigma_j = n(n-1)\overline{r_{ij}\sigma_i\sigma_j}$, as there are $n(n-1)$ covariance terms. This number is simply the number of terms in the $n \times n$ table minus the number of variance terms $= n^2 - n = n(n-1)$. Hence:

$$\sigma_C^2 = \sigma\overline{n}^2 + n(n-1)\overline{r_{ij}\sigma_i\sigma j} \qquad (9:13)$$

If the component scores were in Z score form the σ's would disappear and in Z score terms (9:13) would become:

$$\sigma_C^2 = n + n(n-1)\overline{r_{ij}} \qquad (9:14)$$

(formula in terms of Z score components).

4. *Correlation of a Composite Variable with an Outside Variable*

Recalling the formula for $r_{xy;}r_{xy} = \frac{\sum xy}{N\sigma_x\sigma_y}$, it can be seen that the correlation between a composite variable © and an outside variable (0) will be:

$$r_{0C} = \frac{\sum x_0 x_C}{N\sigma_0\sigma_C} = \frac{\sum x_0(x_1 + x_2 + \ldots x_n)}{N\sigma_0\sigma_C} \qquad (9:15)$$

This equals:

$$\frac{\sum x_0 x_1 + \sum x_0 x_2 + \ldots \sum x_0 x_n}{N\sigma_0\sigma_C}$$

which in turn equals:

$$\frac{\frac{1}{N}\left(\sum x_0 x_1 + \sum x_0 x_2 \ldots \sum x_0 x_n\right)}{\sigma_0\sigma_C}$$

The terms in the numerator are now all covariance terms and the above can therefore be written (following (9:5)), as:

$$\frac{\sigma_0\sigma_1 r_{01} + \sigma_0\sigma_2 r_{02} + \sigma_0\sigma_n r_{0n}}{\sigma_0\sigma_C}$$

Dividing by σ_0 gives:

$$r_{0C} = \frac{\sigma_1 r_{01} + \sigma_2 r_{02} + \ldots \sigma_n r_{0n}}{\sigma_C} \qquad (9{:}16)$$

In Z score terms this becomes (finding σ_C as the square root of (9:13)):

$$r_{0C} = \frac{\sum r_{oi}}{\sqrt{n + n(n-1)\overline{r_{ij}}}} = \frac{n\overline{r_{oi}}}{\sqrt{n + n(n-1)\overline{r_{ij}}}} \qquad (9{:}17)$$

For reasons which will become apparent in a moment it is convenient to divide numerator and denominator by *n* giving:

$$r_{0C} = \frac{\overline{r_{oi}}}{\sqrt{\frac{n}{n^2} + \frac{n(n-1)}{n^2}\overline{r_{ij}}}} = \frac{\overline{r_{oi}}}{\sqrt{\frac{1}{n} + \frac{n-1}{n}\overline{r_{ij}}}} \qquad (9{:}18)$$

As *n* increases in size, i.e. as the number of components increases, the value of $\frac{1}{n}$ becomes smaller and smaller, and $\frac{n-1}{n}$ becomes nearer and nearer to 1. So with a large number of components, as *n* approaches infinity:

$$r_{0C} = \frac{\overline{r_{oi}}}{\sqrt{\overline{r_{ij}}}};\, n \to \infty \qquad (9{:}19)$$

In words the correlation between a composite variable and an outside variable is equal to the mean correlation between components and the outside variable, divided by the square root of the mean intercorrelation between components.

5. *Correlation between two Composite Variables*

Suppose that $C_X = (X_1 + X_2 + ...X_n)$ and $C_Y = (Y_1 + Y_2 + ...Y_n)$,

And that there are *n* components in C_x and *m* components in C_Y.

$$r_{C_X C_Y} \text{ will be } \frac{\sum(x_1 + x_2 + ...x_n)(y_1 + y_2 + ...y_m)}{N\sigma_{C_X}\sigma_{C_y}}$$

By steps which will by now be familiar this becomes firstly:

$$\frac{\sum x_1 y_1 + \sum x_1 y_1 + ...\sum x_n y_m}{N\sigma_{C_X}\sigma_{C_Y}} \qquad (9:20)$$

In turn this becomes:

$$\frac{r_{x_1y_1}\sigma_{x_1}\sigma_{y_1} + r_{x_1y_2}\sigma_{x_1}\sigma_{y_2} + ...r_{x_ny_m}\sigma_{x_n}\sigma_{y_m}}{\sigma_{C_x}\sigma_{C_y}} \qquad (9:21)$$

As $\sum r_{x_iy_i}\sigma_{x_i}\sigma_{y_i} = nm\overline{r_{x_iy_i}\sigma_{x_i}\sigma_{y_i}}$ (9 : 21) becomes :

$$\frac{nm\;\overline{r_{x_iy_i}\sigma_{x_i}\sigma_{y_i}}}{\sqrt{n\overline{\sigma}^2_{x_i} + n(n-1)\overline{r_{x_ix_j}\sigma_{x_i}\sigma_{x_j}}}\sqrt{m\overline{\sigma}^2_{y_i} + m(m-1)\overline{r_{y_iy_j}\sigma_{y_i}\sigma_{y_j}}}} \qquad (9:22)$$

If all components are in *Z* score form, this becomes:

$$\frac{nm\;\overline{r_{x_iy_i}}}{\sqrt{n + n(n-1)\overline{r_{x_iy_i}}}\sqrt{m + m(m-1)\overline{r_{y_iy_j}}}} \qquad (9:23)$$

Dividing numerator and denominator by *nm* gives:

$$\frac{\overline{r_{x_i y_i}}}{\sqrt{\frac{n}{n^2}+\frac{n(n-1)}{n^2}\overline{r_{x_i x_j}}}\sqrt{\frac{m}{m^2}+\frac{m(m-1)}{m^2}\overline{r_{y_i y_j}}}} \tag{9:24}$$

$$=\frac{\overline{r_{x_i y_i}}}{\sqrt{\frac{1}{n}+\frac{n-1}{n}\overline{r_{x_i x_j}}}\sqrt{\frac{1}{m}+\frac{m-1}{m}\overline{r_{y_i y_j}}}}$$

As the number of components in each composite becomes larger $\frac{1}{n}$ and $\frac{1}{m}$ become closer to zero, and $\frac{n-1}{n}$ and $\frac{m-1}{m}$ become nearer to 1. Therefore, as *n* and *m* approach infinity we obtain:

$$\frac{\overline{r_{x_i y_i}}}{\sqrt{\overline{r_{x_i x_j}}}\sqrt{\overline{r_{y_i y_j}}}} \tag{9:25}$$

Problems

A. Suppose a composite is formed of three tests X_1, X_2, and X_3, each with a mean of 10 and a standard deviation of 3. If $r_{x_1x_2}=.30, r_{x_1x_3}=.40,$ and $r_{x_2x_3}=.50,$ what will be the mean and variance of the composite scores?

B. A composite is made up of four tests $X_1, X_2, X_3,$ and X_4, with $M_{x_1}=10,\ M_{x_2}=20,\ M_{x_3}=30, M_{x_4}=40$; and $\sigma_{x_1}=2, \sigma_{x_2}=3, \sigma_{x_3}=4,$ and $\sigma_{x_4}=5.$ If $r_{x_1x_2}=.20, r_{x_1x_3}=.30,$ $r_{x_1x_4}=.30, r_{x_2x_3}=.40,\ r_{x_2x_4}=.40,$ and $r_{x_3x_4}=.20,$ what is the mean of the composite and what is its variance?

C. Given a weighted composite $C=2X_1+3X_2,$ and $r_{12}=.40;$ $M_{x_1}=10,$ and $M_{x_2}=20;$ and $\sigma_{X_1}=5$ and $\sigma_{X_2}=6$, what is the mean of *C* and what is its variance?

D. Given two composites *X* and *Y* what will be their correlation given the following data on the composites:

	X_2	Y_1	Y_2	Y_3	M	σ
X_1	.20	.10	.20	.30	10	3
X_2		.20	.30	.40	15	4
Y_1			.50	.60	20	5
Y_2				.40	25	5
Y_3					30	5

Use Formula (9:21).

Answers

A. $\overline{C} = \overline{X}_1 + \overline{X}_2 + \overline{X}_3 = 10 + 10 + 10 = 30$

$$\sigma_C^2 = \sigma_{X_1}^2 + \sigma_{X_2}^2 + \sigma_{X_3}^2 + 2r_{X_1X_2}\sigma_{X_1}\sigma_{X_2} + 2r_{X_1X_3}\sigma_{X_1}\sigma_{X_3} + 2r_{X_2X_3}\sigma_{X_2}\sigma_{X_3}$$

$$= 9 + 9 + 9 + 2(.30 \times 3 \times 3) + 2(.40 \times 3 \times 3) + 2(.50 \times 3 \times 3)$$

= + 48.6.

B. $\overline{C} = 10 + 20 + 30 + 40 = 100$

$$\sigma_C^2 = 4 + 9 + 16 + 25 + 2(.20 \times 2 \times 3) + 2(.30 \times 2 \times 4)$$
$$+ 2(.30 \times 2 \times 5) + 2(.40 \times 3 \times 4) + 2(.40 \times 3 \times 5) + 2(.20 \times 4 \times 5)$$

= 96.8.

C. $\overline{C} = W_1\overline{X}_1 + W_2\overline{X}_2 = (2 \times 10) + (3 \times 20) = 80$

$$\sigma_C^2 = W_1{}^2\sigma_{X_1}^2 + W_2{}^2\sigma_{X_2}^2 + 2W_1W_2r_{X_1X_2}\sigma_{X_1}\sigma_{X_2}$$

$$= (4 \times 25) + (9 \times 36) + (2 \times 2 \times 3 \times .40 \times 5 \times 6)$$

$$= 568.$$

D.

$$r_{C_xC_y} = \frac{\begin{array}{c} r_{X_1Y_1}\sigma_{X_1}\sigma_{Y_1} + r_{X_1Y_2}\sigma_{X_1}\sigma_{Y_2} + r_{X_1Y_3}\sigma_{X_1}\sigma_{Y_3} \\ +r_{X_2Y_1}\sigma_{X_2}\sigma_{Y_1} + r_{X_2Y_2}\sigma_{X_2}\sigma_{Y_2} + r_{X_2X_3}\sigma_{X_2}\sigma_{Y_3} \end{array}}{\sigma_{C_X}\sigma_{C_Y}}$$

The numerator = $(.10 \times 3 \times 5) + (.20 \times 3 \times 5)$
$+(.30 \times 3 \times 5) + (.20 \times 4 \times 5)$
$+(.30 \times 4 \times 5) + (.40 \times 4 \times 5)$
= 27.0

The denominator =

$$\sqrt{\sigma_{C_X}^2} = \sqrt{9 + 16 + 2(.20 \times 3 \times 4)} = \sqrt{29.8}$$

$$\sqrt{\sigma_{C_Y}^2} = \sqrt{\begin{array}{c} 25 + 25 + 25 + 2(.50 \times 5 \times 5) + 2(.60 \times 5 \times 5) \\ + 2(.40 \times 5 \times 5) \end{array}}$$

$$= \sqrt{150}$$

Therefore $r_{C_XC_Y} = \dfrac{27.0}{\sqrt{29.8}\sqrt{150}} = .40$

6. *The Multiple Regression Equation Again*

On reflection it will by now be clear that in multiple regression, the regression weights combined appropriately with the raw scores or *Z* scores yield weighted composite scores. These scores when correlated with the criterion scores result in the multiple correlation coefficient. Thus $R_{0.12...n}$, although it expresses the relationship between a number of variables and the criterion, is in fact the correlation between only two variables - the weighted composite variable and the criterion.

Thus the mean score on the predicted variable and its variance can be worked out from the formulae for weighted composites. It is merely necessary to substitute β's or b's for the w's in Formulae 9:2 and 9:12. It can be shown that the mean of the composite will equal the mean of the criterion.

$$\hat{M}_0 = M_0 \qquad (9:26)$$

Where:

$\hat{M}_0 =$ the mean of the predicted scores, and
$M_0 =$ the mean of the criterion scores

Proof

(1) $\hat{X}_0 = M_0 - b_1M_1 - ...b_nM_n + b_1X_1 + ...b_nX_n$

(2) $\hat{M}_0 = \dfrac{\sum(M_0 - b_1M_1 - ...b_nM_n + b_1X_1 + ...b_nX_n)}{N}$

(3) $= \dfrac{NM_0 - b_1NM_1 - b_nNM_n + b_1\sum X_1 + ...b_n\sum X_n}{N}$

(4) $= M_0 - b_1M_1 - ...b_nM_n + b_1M_1 + ...b_nM_n$

(5) $= M_0$

The variance of the predicted scores will be:

$$\hat{\sigma}_0^2 = \sum b_i^2\sigma_i^2 + 2\sum b_i b_j \sigma_i \sigma_j r_{ij} \tag{9:27}$$

Proof

Once more a table will help in working out the products of the weighted deviation scores:

	b_1x_1	b_2x_2	...	b_nx_n
b_1x_1	$b_1{}^2x_1{}^2$	$b_1x_1b_2x_2$	...	$b_1x_1b_nx_n$
b_2x_2	$b_1x_1b_2x_2$	$b_2{}^2x_2{}^2$	...	$b_2x_2b_nx_n$
...	...	...	...	...
b_nx_n	$b_1x_1b_nx_n$	$b_2x_2b_nx_n$	...	$b_n{}^2x_n{}^2$

Summing these values across individuals will give:

(1) $\sum \hat{x}_0^2 = b_1^2\sum x_1^2 + b_2^2\sum x_2^2 + ..b_n^2\sum x_n^2 + 2b_1b_2\sum x_1x_2 + ...2b_{(n-1)}b_n\sum x_{(n-1)}x_n$

(2) Dividing by *N* gives:

$$b_1^2\sigma_1^2 + b_2^2\sigma_2^2 + ..b_n^2\sigma_n^2 + 2b_1b_2\sigma_1\sigma_2r_{12} + ...2b_{(n-1)}b_n\sigma_{(n-1)n}$$

(3) This equals:

$$\sum b_i^2\sigma_i^2 + 2\sum b_i b_j \sigma_i \sigma_j r_{ij}$$

Problems

A. If $r_{01} = .30, \quad r_{02} - .40, \text{ and } \quad r_{12} = .50,$ write the *Z* score regression equation for predicting 0 from 1 and 2.

B. What will be the mean of the predicted scores? (Use the formula for the mean of a weighted composite.)

C. What will the variance of the predicted scores be? (Use the formula for the variance of a weighted composite.)

Answers

A. $$Z_0 = \beta_1 Z_1 + \beta_2 Z_2 ; \beta_1 = \frac{r_{01} - r_{02} r_{12}}{1 - r_{12}^2}$$

and $\beta_2 = \frac{r_{02} - r_{01} r_{12}}{1 - r_{12}^2}$

Therefore $\beta_1 = \frac{.30 - (.40 \times .50)}{1 - .50^2} = .13$

$$\beta_2 = \frac{.40 - (.30 \times .50)}{1 - .50^2} = .33$$

Therefore $Z_0 = .13 Z_1 + .33 Z_2$

B. $$\hat{M}_{Z_0} = .13 M_{Z_1} + .33 M_{Z_2} = 0$$

C. (1) A square table will help:

	$\beta_1 Z_1$	$\beta_2 Z_2$
$\beta_1 Z_1$	$\beta_1^2 Z_1^2$	$\beta_1 \beta_2 Z_1 Z_2$
$\beta_2 Z_2$	$\beta_2 \beta_1 Z_2 Z_1$	$\beta_2^2 Z_2^2$

(2) Summing across individuals and dividing by *N* gives:

$$\hat{\sigma}_{Z_0}^2 = \beta_1^2 \frac{\sum Z_1^2}{N} + \beta_2^2 \frac{\sum Z_2^2}{N} + 2\beta_1\beta_2 \frac{\sum Z_1 Z_2}{N}$$

(3) As $\frac{\sum Z_1^2}{N}$ and $\frac{\sum Z_2^2}{N}$ equal 1; and as $\frac{\sum Z_1 Z_2}{N}$ equals r_{12} we obtain:

(4) $\hat{\sigma}_{Z_0}^2 = \beta_1^2 + \beta_2^2 + 2\beta_1\beta_2 r_{12}$

= $.13^2 + .33^2 + (2 \times .13 \times .33 \times .50)$

= .17

(5) Therefore $\hat{\sigma}_{Z_0} = .41$.

According to (8:13) $\hat{\sigma}_0 = R_{0.12}\sigma_0$.

Does this agree with the answer just obtained?

Item statistics

1. *The Mean of Dichotomous Item*

In this chapter we will be concerned with dichotomous items, i.e. items cored either 1 or 0. Such items might be individual items in an intelligence test, or questions on a personality inventory. As the formula for the mean is $\sum X/N$, and as dichotomous items, (hereafter called simply 'items'), are only scored 1 or 0.

$$M = \frac{\sum X}{N} = \frac{\sum 1's + \sum 0's}{N} = \frac{\sum 1's}{N} \qquad (10:1)$$

However by Summation Rule 2 the sum of a constant taken n times is n times that constant, so (10:1) becomes, as the constant is 1,

$$M_{(item)} = \frac{n}{N} \qquad (10:2)$$

So the mean of a dichotomous item is the number of cases scoring 1 divided by the total number of cases. Further the number of cases scoring 1 divided by the total number of cases equals the proportion of cases scoring 1. Hence, using 'p' as the symbol for a proportion:

$$M_{(item)} = \frac{n}{N} = p \qquad (10:3)$$

It will be convenient to also have a symbol for the proportion not passing an item, and this we will call 'q'. Because the total proportion must be 1, (10:4) is obtained.

$$q = 1 - p \qquad (10:4)$$

Problems

A. If 10 out of 50 subjects pass an item, what is mean score for that item?

B. If 19 out of 25 subjects score 1 on a dichotomous item what is its mean?

C. What is the value of *q* in (A) and in (B)?

Answers

A. 20;

B. .76;

C. (A) = .80, (B) = .24.

2. *The Variance of a Dichotomous Item*

Amongst the formulae for the variance described earlier was $(2:7), \sigma^2 = \frac{\sum X^2}{N} - M^2$. In the case where all the *X*'s are 1's or 0's, $\sum X$ will also equal *n*.

$$\sum X_{(item)} = \sum X^2_{(item)} = n \qquad (10:5)$$

Substituting these values in formula (2:7) for the variance and using (10:3) gives

$$\sigma^2_{(item)} = \frac{n}{N} - \left(\frac{n}{N}\right)^2 = p - p^2 \qquad (10:6)$$

This can be written as:

$$\sigma^2_{(item)} = p - p^2 = p(1-p) = pq. \qquad (10:7)$$

(from (10:4) 1 - *p* = *q*).

Thus the variance of an item is *pq.*

Problems

A. If 20 out of 50 people pass an item what is its variance?

B. If 60 out of 200 people score 1 on a dichotomous item what is its variance?

C. What is the maximum variance that an item can have?

Answers

A. .24;

B. .21;

C. .25; i.e. when $p = q = .50$. (If in doubt prepare a table with values of p of .10, .20, .30, .40, .50, .60, .70, .80, .90 and work out the value of pq).

3. *The Covariance of Items*

It will be recalled that one formula for the covariance is

$$S_{xy} = \frac{\sum XY}{N} - M_x M_y \qquad (9:6)$$

Letting two items be called *I* and *J*. $M_i M_j$ becomes $p_i p_j$ (using (10:3)). To work out the value of $\sum IJ$ consider the following table.

		Item J	
		0	1
Item	0	0	0
1	1	0	1

In the body of the table are the products of all possible pairs of scores on the two items. Only in the case where the individual passes both items will the product of his item scores have a non-zero value. So the value of $\sum IJ$ in terms of items will equal the number of subjects passing both items, and $\frac{\sum IJ}{N}$ will be the proportion of subjects passing both items. The symbol used for the proportion passing both items will be p_{ij}, so the formula for the covariance of items is:

$$S_{ij} = p_{ij} - p_i p_j \qquad (10:8)$$

Problems

If, in a group of 50 subjects 20 pass item *X* and 40 pass item *Y*, and 18 pass items *X* and *Y*,

A. What is the mean for item *X*?

B. What is the mean for item *Y*?

C. What is the variance of item *X*?

D. What is the variance of item *Y*?

E. What is the covariance of items *X* and *Y*?

Answers

A. .40;

B. .80;

C. .24;

D. .16;

E. .04.

4. *The Mean of a Combination of Items*

A composite variable composed of *m* items will have a mean equal to:

$$\overline{C}_{(item)} = p_1 + p_2 + p_3 + ...p_m \qquad (10:9)$$

where $p_1 p_2 ... p_m$ are the proportions passing items 1, 2 and *m*.

Proof

(1) $\overline{C}_{(item)} = \dfrac{\sum C_{(item)}}{N}$

(2) $\dfrac{\sum C_{(item)}}{N} = \dfrac{n_1 + n_2 + n_3 + ...n_m}{N}$

(3) $= \dfrac{n_1}{N} + \dfrac{n_2}{N} + \dfrac{n_3}{N} + ...\dfrac{n_m}{N}$

(4) $= p_1 + p_2 + p_3 + ...p_m$

5. *The Variance of a Combination of Items*

In Chapter 9 it was shown that the variance of a composite is equal to

$$\sum \sigma_i^2 + n(n-1)\overline{r_{ij}\sigma_i\sigma_j}$$

Further from (9:5) it is known that $r_{ij}\sigma_i\sigma_j$ equals $\frac{\sum ij}{N}$ giving

$$\sigma_C^2 = \sum \sigma_1^2 + n(n-1)\overline{\frac{\sum ij}{N}} \qquad (10:10)$$

In terms of dichotomous items $\sigma_2 = pq;$ and $S_{ij} = p_{ij} - p_i p_j$, so (10:10) becomes

$$\sigma_C^2 = \sum p_i q_i + n(n-1)\left(\overline{p_{ij} - p_i p_j}\right) \qquad (10:11)$$

or $$\sigma_C^2 = \sum p_i q_i + 2\sum \left(p_{ij} - p_i p_j\right) \qquad (10:12)$$

Problems

Given the following data:

		Item		
Subject	*A*	B	C	Y
1	1	0	1	1
2	0	0	1	2
3	1	0	1	3
4	0	0	1	4
5	1	0	1	5
6	0	1	1	1
7	1	1	1	2
8	0	1	1	3
9	1	1	1	4
10	0	1	1	5

A. What are the means of items *A, B* and *C*?

B. What are the variances of items *A, B* and *C*?

C. What are the values of covariance *AB*, covariance *AC* and covariance *BC*?

D. What is the mean of the composite (*A* + *B* + *C*)?

E. What is the variance of the composite (*A* + *B* + *C*)?

F. What are the variances of the composites (*A* + *B*), (*A* + *C*), and (*B*+*C*)?

G. Check your answers by forming the appropriate composites.

Answers

A. The mean of an item is $\frac{n}{N} = p$.

$$\overline{A} = .50; \quad \overline{B} = .50; \quad \overline{C} = 1.0.$$

B. The variance of an item is *pq*.

$$\sigma_a^2 = .25; \quad \sigma_b^2 = .25; \quad \sigma_c^2 = 0.$$

C. $S_{ab} = .20 - (.50 \times .50) = -.50; \quad S_c^a = .50 - (1.0 \times .50) = 0$
$S_{bc} = .50 - (1.0 \times x.50) = 0.$

D. $\overline{C} = \overline{A} + \overline{B} + \overline{C} = 2.0$

E. $\sigma_c^2 = \sigma_a^2 + \sigma_b^2 + \sigma_c^2 + 2(S_{ab} + S_{ac} + S_{bc})$
$= .25 + .25 + 0 + 2(-.05 + 0 + 0) = .40$

F.

$$\sigma_{ab}^2 = \sigma_a^2 + \sigma_b^2 + 2s_{ab} = .50 + 2(-.05) = .40$$

$$\sigma_{ac}^2 = \sigma_a^2 + \sigma_c^2 + 2s_{ac} = .50 + 0 + 0 = .25$$

$$\sigma_{bc}^2 = \sigma_b^2 + \sigma_c^2 + 2s_{bc} = .50 + 0 + 0 = .25$$

6. *The Correlation between One Dichotomous Item and Another*

One formula for the product moment correlation coefficient is:

The covariance divided by the product of the standard deviation of the two variables, i.e.

$$r_{xy} = \frac{\sum xy / N}{\sigma_x \sigma_y} \qquad (10{:}13)$$

In terms of items this becomes:

$$r_{ij} = \frac{p_{ij} - p_i p_j}{\sqrt{p_i q_i}\sqrt{p_j q_j}} \qquad (10{:}14)$$

This is the formula for the Phi Coefficient. For computational purposes a slightly different formula is used. Consider the following table:

		Item J		
		1	0	
Item I	1	*a*	*b*	*a* + *b*
	0	*c*	*d*	*c* + *d*
		a + *c*	*b* + *d*	*N*

The covariance will equal;

$$\frac{a}{N} - \left(\frac{a+b}{N}\right)\left(\frac{a+c}{N}\right)$$

The variance of *I* will equal:

$$\left(\frac{a+b}{N}\right)\left(\frac{c+d}{N}\right)$$

and its square root will equal the standard deviation of *I.*

$$\left(\frac{a+c}{N}\right)\left(\frac{b+d}{N}\right)$$

So the formula becomes:

$$\frac{\frac{a}{N}-\left(\frac{a+b}{N}\right)\left(\frac{a+c}{N}\right)}{\sqrt{\left(\frac{a+b}{N}\right)\left(\frac{c+d}{N}\right)}\sqrt{\left(\frac{a+c}{N}\right)\left(\frac{b+d}{N}\right)}}$$

which can be simplified to:

$$\frac{ad-bc}{\sqrt{(a+b)(c+d)(a+b)(b+d)}} \qquad (10:15)$$

Proof

$$1=\frac{\frac{a}{N}-\frac{(a+b)(a+c)}{N^2}}{\sqrt{\frac{(a+b)(c+d)}{N^2}}\sqrt{\frac{(a+c)(b+d)}{N^2}}} \qquad (10:16)$$

(2) The denominator can be transformed by taking N^2 from under the root sign to:

$$\frac{1}{N}\sqrt{(a+b)(c+d)}.\frac{1}{N}\sqrt{(a+c)(b+d)}$$
$$=\frac{1}{N^2}\sqrt{(a+c)(c+d)(a+c)(b+d)}$$

(3) Multiplying numerator and denominator by N^2 gives

$$\frac{Na-(a+b)(a+c)}{\sqrt{(a+c)(c+d)(a+c)(b+d)}}$$

(4) As $N = a + b + c + d$ the numerator can be simplified to:

$$a(a+b+c+d)-(a+b)(a+c)$$

(5) Multiplying appropriately this becomes:

$$a^2+ab+ac+ad-a^2-ac-ab-bc$$

(6) which equals $ad-bc$

(7) Thus:

$$r_{ij}=\frac{ad-bc}{\sqrt{(a+c)(c+d)(a+c)(b+d)}}$$

7. *The Correlation between a Dichotomous Item and a Continuous Variable*

Again using formula (10:13) and substituting (9:6) for the covariance, it is possible to work out the product moment correlation between a dichotomous item and a continuous variable. In this case the covariance will consist of the mean of scores on the continuous variable of those who obtained a score of 1 on the dichotomous item. For those who scored 0 on the item the product of the item and variable score will of course be zero. Where *I* is the item and *Y* the continuous variable:

$$r_{ij} = \frac{\sum Y_{(1)} / N - p_i M_y}{\sigma_y \sqrt{p_i q_i}} \qquad (10:17)$$

Where $\sum Y_{(1)}$ is the sum of *Y* scores for those scoring 1 on the item and $p_i = M_i$ and $\sqrt{p_i q_i} = \sigma_i$

(10:17) is called the Point-Biserial Coefficient. Its value is worked out as follows:

Step 1. Find the sum of *Y* scores for subjects scoring 1 on the item.

Step 2. Divide the value found in Step 1 by *N*.

Step 3. Find M_i and *y*.

Step 4. Find the product M_iM_y.

Step 5. Subtract the value in Step 4 from the value in Step 2.

Step 6. Find σ_i and σ_y.

Step 7. Find the product $\sigma_i\sigma_y$.

Step 8. Divide the value found in Step 5 by the value found in Step 7.

Problems

Given the data in the problems at the end of Section 10:5:

A. What are the values of:

(1) r_{ab} ?
(2) r_{ac} ?
(3) r_{bc} ?

B. What are the values of:

(1) r_{ay} ?
(2) r_{by} ?
(3) r_{cy} ?

Answers

A. (1) - .20;
(2) .00;
(3) .00.

B. (1) .00;
(2) .00;
(3) .00.

Chapter Eleven

Reliability

1. *Introduction*

Psychological measurement is more prone to error than physical measurement, and reliability is best defined in terms of error. When a measurement contains much error it is said to have low reliability, or to be unreliable, while if it contains little error it is said to be reliable or to have high reliability. Because of the susceptibility to error of psychological measurements, psychologists have always been interested in developing theories of measurement error. The present chapter aims to introduce the reader to some of the theory of reliability and to the various ways of assessing the reliability of a measuring device. At this stage it is desirable not to be too specific about the definition of error. A classification of different types of error will be presented later, but for present purposes 'error' is best left as an abstract term.

As far as theory is concerned, the two theories of measurement error discussed will be:

(a) the theory of parallel tests and true and error scores,

(b) the theory of domain sampling.

2. Parallel Tests, True Scores, and Error Scores

Basic to the theory of true scores and error scores is the assumption that an obtained score is made up of two components:

(a) the individual's true score

(b) the individual's error score, which can be positive or negative.

This can be written as an equation.

$$X = T + E \tag{11:1}$$

Where X = the obtained score
T = the true score
E = the error score

An individual's true score can be conceptualized in this theory as almost a metaphysical platonic idea. It is the individual's real score. Suppose someone was tested on an intelligence test, it might well be asked whether the result represented his real intelligence. Anyone asking such a question is implicitly accepting the notion of a true score as a real characteristic of an individual. Other conceptions of true score are possible and in some ways preferable, but for introductory purposes the notion of platonic true scores is easier to deal with and a grasp of measurement theory using platonic true scores is easily transferred to other theoretical models.

A number of further assumptions are also made:

Assumption 1: the action of error is completely random.

Assumption 2: the mean error score is zero.

Assumption 3: the variance of error scores is the same on all tests of a given length.

Assumption 4: the correlations between error scores and:

(a) true scores
(b) other error scores,

are zero.

Any of these really follow from assuming that error will be random in its effects. For example random effects would be just as likely to increase as to decrease a score, therefore Assumption 2 is reasonable. If error is random it is reasonable to assume that error variance on different tests will be the same and so on.

Now suppose that there are a number of tests of a given length which measure the same trait or traits to the same degree. Such tests are called parallel tests. Given a number of such tests it is possible to deduce certain of their properties.

3. *The Means of Parallel Tests*

Firstly the mean score on any of the tests given to a large number of individuals will be the mean of the true scores:

$$M_1 = M_2 = ...M_n = \overline{T} \qquad (11:2)$$

Where $M_1, M_2,$ and M_n are the means of tests 1,2, and *n* respectively and $\overline{T}$ equals the mean true score.

Proof

(1) $M_1 = \frac{\sum X_1}{N}$

(2) $\frac{\sum X_1}{N} = \frac{\sum (T+E)}{N}$

(3) $\frac{\sum (T+E)}{N} = \frac{\sum T}{N} + \frac{\sum E}{N}$

(4) $\frac{\sum T}{N} + \frac{\sum E}{N} = \overline{T} + \overline{E}$

(5) By Assumption 2 above $\overline{E} = 0.$ So

$$\mathrm{M}_1 = \overline{T} + \overline{E} = \overline{T}$$

This is true of any parallel test, so:

(6) $M_1 = M_2 = ...M_n = \overline{T}.$

4. *The Variances of Parallel Tests*

A second deduction is that the variances of the tests will be equal, but before proving this it will be shown that

$$x = t + e \qquad (11:3)$$

where $x = X - M$

$t = T - \overline{T}$

$e = E - \overline{E}$

Proof

(1) $x = X - M$

(2) $X - M = (T + E) - \frac{\sum(T + E)}{N}$

(3) $(T + E) - \frac{\sum(T + E)}{N} = (T + E) - (\overline{T} + \overline{E})$

(4) $(T + E) - (\overline{T} + \overline{E}) = (T - \overline{T}) + (E - \overline{E})$

(5) But $T - \overline{T} = t$ and $E - \overline{E} = e$

So

$$x = t + e$$

It will now be fairly easy to show that the variances of the tests will equal one another.

$$\sigma_i^2 = \sigma_1^2 = \sigma_2^2 = \ldots\sigma_n^2 \qquad (11:4)$$

Proof

(1) $\sigma_i^2 = \frac{\sum x^2}{N}$

(2) $\frac{\sum x^2}{N} = \frac{\sum (t+e)^2}{N}$ $(\text{using}(11.3))$

(3) $\frac{\sum (t+e)^2}{N} = \frac{\sum (t^2 + e^2 + 2et)}{N}$

(4) $= \frac{\sum t^2}{N} + \frac{\sum e^2}{N} + 2\frac{\sum et}{N}$

(5) But $\frac{\sum et}{N}$ is a covariance and equals $r_{et}\sigma_e\sigma_t \frac{\sum x^2}{N} = \frac{\sum (t+e)^2}{N}$

By Assumption 4 $r_{et} = 0$ s0 (3) becomes:

$$\frac{\sum (t+e)^2}{N} = \sigma_t^2 + \sigma_e^2$$

(6) By Assumption 3, σ_e^2 is the same for all tests, and the variance of true scores must always be the same, so

$$\sigma_i^2 = \sigma_t^2 + \sigma_e^2 = \sigma_1^2 = \sigma_2^2 = \sigma_n^2$$

5. *The Intercorrelation of Parallel Tests*

So far it has been shown that parallel tests have the same mean and the same variance. It will now be shown that the correlation between any two parallel tests equals the correlation between any other two.

$$r_{ij} = r_{12} = r_{13} = \ldots r_{(n-1)n} \qquad (11:5)$$

Proof

(1) $r_{ij} = \dfrac{\sum x_i x_j}{N\sigma_i\sigma_j} = \dfrac{\sum(t+e_i)(t+e_j)}{N\sigma_i\sigma_j}$

(3) $\dfrac{\sum t^2 + \sum e_i e_j + \sum e_i t + \sum e_j t}{N\sigma_i\sigma_j}$

(4) Dividing numerator and denominator by *N* gives:

$$\frac{\sum t^2/N + \sum e_i e_j/N + \sum e_i t/N + \sum e_j t/N}{\sigma_i\sigma_j}$$

(5) The numerator consists of a variance term and a number of covariance terms so: (4) becomes

$$\frac{\sigma_t^2 + r_{e_i e_j}\sigma_{e_i}\sigma_{e_j} + r_{e_i t}\sigma_t + r_{e_j t}\sigma_{e_j}\sigma_t}{\sigma_i\sigma_j}$$

(6) The last three terms in the numerator equal zero, and it has been shown in (11:4) that $\sigma_i = \sigma_j$ so:

$$r_{ij} = \frac{\sigma_t^2}{\sigma^2}$$

6. *The Reliability Coefficient*

The correlation between any two parallel tests is equal to the proportion of variance which is true variance. The value obtained by correlating two parallel tests is called the reliability coefficient, symbolised as r_{xx}.

$$r_{xx} = \frac{\sigma_t^2}{\sigma_x^2} \text{ = reliability coefficient of test } X \qquad (11:6)$$

Below are listed some other equations involving r_{xx}. In deriving these it is helpful to recall from (11:4), Step 5, that: $\sigma_x^2 = \sigma_t^2 + \sigma_e^2$

$$r_{xx}\sigma_x^2 = \sigma_t^2 \qquad (11:7)$$

(11:7) is obtained from (11:6) by multiplying both sides by σ_x^2. To obtain the variance of true scores it is necessary to multiply the variance of obtained scores by the reliability coefficient. In practice r_{xx} is always less than 1.0 so σ_t^2 is always less than 1.0 so σ_t^2 will be less than σ_x^2.

$$r_{xx} = 1 - \frac{\sigma_e^2}{\sigma_x^2} \qquad (11:8)$$

Proof

(1) $r_{xx} = \frac{\sigma_t^2}{\sigma_x^2}$ and

(2) $\sigma_x^2 = \sigma_t^2 + \sigma_e^2$ so

(3) $\sigma_x^2 - \sigma_e^2 = \sigma_t^2$ so

(4) $r_{xx} = \frac{\sigma_x^2 - \sigma_e^2}{\sigma_x^2} = 1 - \frac{\sigma_e^2}{\sigma_x^2}$

The variance of error scores is also derivable quite easily and this is of great value for a number of purposes.

$$\sigma_e^2 = \sigma_x^2(1 - r_{xx}) \qquad (11{:}9)$$

Proof

(1) $r_{xx} = 1 - \dfrac{\sigma_e^2}{\sigma_x^2}$

(2) $r_{xx} + \dfrac{\sigma_e^2}{\sigma_x^2} = 1$

(3) $\dfrac{\sigma_e^2}{\sigma_x^2} = 1 - r_{xx}$

(4) (Multiplying both sides by σ_x^2)

$$\sigma_e^2 = \sigma_x^2(1 - r_{xx})$$

The square root of this value is called the standard error of measurement, and it represents the standard deviation of the distribution of obtained scores about true scores. This will be discussed in more detail later.

7. *The Correlation of Parallel Tests with an Outside Variable*

Another feature of parallel tests is that the correlation of any one of them with an outside variable is the same as the correlation of any other with the outside variable.

Calling the parallel tests $X_1, X_2, \ldots X_n$ and the outside variable *Y:*

$$r_{x_i y} = r_{x_1 y} = \ldots r_{x_n y} \qquad (11:10)$$

Proof

(1) $r_{x_i y} = \dfrac{\sum x_i y}{N\sigma_{x_i}\sigma_y}$

(2) $\dfrac{\sum x_i y}{N\sigma_{x_i}\sigma_y} = \dfrac{\sum(t+e)y}{N\sigma_{x_i}\sigma_y}$

(3) So $r_{x_i y} = \dfrac{\sum ty + \sum ey}{N\sigma_{x_i}\sigma_y}$

(4) Dividing numerator and denominator by *N* gives:

$$\frac{\sum ty/N + \sum ey/N}{\sigma_{x_i}\sigma_y}$$

(5) $\sum ey/N = r_{ey}\sigma_e\sigma_y = 0$ and $\sum ty/N = r_{ty}\sigma_t\sigma_y$

So: $r_{x_i y} = \dfrac{r_{ty}\sigma_t\sigma_y}{\sigma_{x_i}\sigma_y}$

(6) Dividing numerator and denominator by σ_y gives:

$$r_{x_i y} = \frac{r_{ty}\sigma_t}{\sigma_{x_i}} = r_{ty}\frac{\sigma_t}{\sigma_{x_i}}$$

(7) $\dfrac{\sigma_t}{\sigma_{x_i}} = \sqrt{\dfrac{\sigma_t^2}{\sigma_{x_i}^2}} = \sqrt{r_{xx}}$

(8) Therefore $r_{x_i y} = r_{ty}\sqrt{r_{xx}}$

Because the true scores are the same on all parallel tests r_{ty} will be the same for all tests and r_{xx} is the reliability coefficient, therefore the correlation between any parallel test and an outside criterion will be found by:

$$r_{x_i y} = r_{ty}\sqrt{r_{xx}} \qquad (11:11)$$

A brief summary may be useful at this point:

(1) Parallel tests have:

(a) the same mean
(b) the same variance
(c) the same correlation between any pair of them
(d) the same correlation with an outside variable.

(2) Two useful measures of reliability are:

(a) the reliability coefficient r_{xx} which is the ratio of true variance to total variance: and
(b) the standard error of measurement which is the Standard deviation of the distribution of obtained scores about true scores.

It will have been noted that r_{xx}, the reliability coefficient, is equal to the proportion of variance in obtained scores accountable for by true scores. It is therefore a coefficient of determination. The square root will be the correlation between true and obtained scores, and is called the index of reliability.

$$r_{x_i t} = \sqrt{\frac{\sigma_t^2}{\sigma_x^2}} = \sqrt{r_{xx}} \qquad (11:12)$$

Again it is emphasised that the reliability coefficient r_{xx} is *not* the correlation between true and obtained scores.

8. *The Length of Parallel Tests and Reliability*

Finally, the relationship between the length of the parallel tests and its effect on the reliability coefficient is worth examining. Suppose that a parallel test is doubled in length by adding to it another parallel test. The combination, being a parallel composite of two tests, is now correlated with another parallel composite of two tests. So in all there are four parallel tests arranged in two pairs. If one parallel composite C_x is made up of tests Y_1 and Y_2.

$$r_{C_xC_y} = \frac{2r_{xx}}{1 + r_{xx}} = \text{reliability of a test double in length} \qquad (11:13)$$

Proof

(1)
$$r_{C_xC_y} = \frac{\sum (x_1 + x_2)(y_1 + y_2)}{N\sqrt{\frac{\sum (x_1 + x_2)^2}{N}}\sqrt{\frac{\sum (y_1 + y_2)^2}{N}}}$$

(2)
$$= \frac{\sum x_1y_1 + \sum x_1y_2 + \sum x_2y_1 + \sum x_2y_2}{N\sqrt{\frac{\sum \left(x_1^2 + x_2^2 + 2x_1x_2\right)}{N}}\sqrt{\frac{\sum \left(y_1^2 + y_2^2 + 2y_1y_2\right)}{N}}}$$

(3) Dividing numerator and denominator by N and remembering that $\frac{\sum xy}{N} = r_{xy}\sigma_x\sigma_y$ gives

$$\frac{r_{x_1y_1}\sigma_{x_1}\sigma_{y_1} + r_{x_1y_2}\sigma_{x_1}\sigma_{y_2} + r_{x_2y_2}\sigma_{x_2}\sigma_{y_1} + r_{x_2y_2}\sigma_{x_2}\sigma_{y_2}}{\sqrt{\sigma_{x_1}^2 + \sigma_{x_2}^2 + 2r_{x_1x_2}\sigma_{x_1}\sigma_{x_1}}\sqrt{\sigma_{y_1}^2 + \sigma_{y_2}^2 + 2r_{y_1y_2}\sigma_{y_1}\sigma_{y_2}}}$$

(4) But the components X_1, X_2, Y_1, Y_2 are all parallel tests to be equal ((11:4) and (11:5)); so (3) becomes:

$$\frac{r_{xx}\sigma_x^2 + r_{xx}\sigma_x^2 + r_{xx}\sigma_x^2 + r_{xx}\sigma_x^2}{\sigma_x^2 + \sigma_x^2 + 2r_{xx}\sigma_x^2}$$

(5) Dividing numerator and denominator by σ_x^2 gives:

$$\frac{4r_{xx}}{2+2r_{xx}}$$

(6) Dividing by 2 gives:

$$\frac{2r_{xx}}{1+r_{xx}}$$

If there had been parallel composites consisting of four components then in step (4) above there would have been 16 terms in the numerator of the type $r_{xx}\sigma_x^2$, 4 variance terms in the denominator, and $12 = n(n-1)$ covariance terms also in the denominator. (If this is not clear use the tabular method suggested in Chapter 9 for working out the variance of a composite.) So in the case of a test 4 times as long:

$$r_{C_xC_y} = \frac{4r_{xx}}{1+3r_{xx}} \qquad (11:14)$$

If the test had been *k* times as long there would have been *k*2 terms of the type $r_{xx}\sigma_x^2$ in the numerator, and the denominator would have also consisted of *k*2 terms, *k* of which would have been variances and *k*(*k* - 1) of which would have been covariances. So a general formula for the reliability of a test lengthened *k* times is:

$$r_{c_xc_y} = \frac{kr_{xx}}{1+(k-r)r_{xx}} \qquad (11:15)$$

This formula is called the Spearman-Brown formula, and it will be reached by a slightly different route in a later section in connection with the theory of domain sampling, which will now be discussed.

9. *The Domain Sampling Model*

In the domain sampling model a test is thought of as being a random sample of all possible items relevant to the characteristic which the test measures. This universe of all possible items from which the sample is drawn is called a domain. In terms of this model an individual's true score is the score he would get if all items in the domain were administered to him.

The model makes a major assumption, which is that the average intercorrelation of an item with all other items is the same for all items. As the number of items in a domain tends to be infinitely large this does not seem an unreasonable assumption. Supposing that there are *n* items in the domain the correlation matrix for the correlations between items can be represented thus:

$$\begin{array}{ll|llll}
 & & 1 & 2 & \ldots & n \\
\hline
\textit{Item} & 1 & r_{11} & r_{12} & \ldots & r_{1n} \\
 & 2 & r_{21} & r_{22} & \ldots & r_{2n} \\
 & . & . & . & \ldots & . \\
 & . & . & . & \ldots & . \\
 & . & . & . & \ldots & . \\
 & n & r_{n1} & r_{n2} & \ldots & r_{nn}
\end{array}$$

The assumption states that the sum of any row in the matrix is equal to the sum of any other row, and also equal to the sum of any column. Dividing either a row or column total by *n* will give the mean intercorrelation between a given item and all other items. Because the mean intercorrelation of an item with all other items is the same for any item, it follows that the mean intercorrelation of all of the items in the domain, will also be equal to the mean correlation of an item with all others.

These assumptions can be stated as follows:

$$\overline{r_{1j}} = \overline{r_{2j}} = \ldots \overline{r_{nj}} = \overline{r_{ij}} \qquad (11:16)$$

Where $\overline{r_{1j}}$ = the average correlation of the first item with other items.

$\overline{r_{2j}}$ = the average correlation of the second item with other items.

$\overline{r_{1j}}$ = the mean correlation between items for the whole matrix.

To prove that $\overline{r_{1j}}$, etc = $\overline{r_{1j}}$ let *R* stand for the sum of any row in the matrix.

(1) $\overline{r_{1j}} = \dfrac{R_1}{n}$; $\overline{r_{2j}} = \dfrac{R_2}{n}$; $\overline{r_{nj}} = \dfrac{R_n}{n}$

(2) Summing the row totals gives the grand total

$$\sum R = R_1 + R_2 + \ldots R_n$$

(3) But by assumption $R_1 = R_2 = \ldots R_n$ so R_n is constant, therefore:

$$\sum R = nR$$

(4) To obtain the mean correlation for the whole matrix the grand total needs to be divided by the total number of correlations in the matrix, i.e. n^2, therefore::

$$\overline{r_{ij}} = \frac{nR}{n^2} = \frac{R}{n}$$

So the mean of all correlations equals the mean of the correlation of any item with all other items.

10. *The Correlation of an Item with the True Score*

In some of the following derivations it is necessary to recall that $\frac{\sum xy}{N\sigma_x\sigma_y} = \frac{\sum Z_x Z_y}{N(1)(1)}$; and further that if the items making up a composite are in *Z* score form:

(a) $\overline{Z}_c = 0$

(b) $(Z_1 + Z_2 + \ldots Z_n) - \overline{Z_c} = (Z_1 + Z_2 + \ldots Z_n)$

(c) $\sigma^2_{Z_c} = \frac{\sum (Z_1 + Z_2 + \ldots Z_n)^2}{N}$

Bearing the above in mind it can be shown that the correlation between an item and the true score is:

$$r_{1t} = \sqrt{\overline{r_{1j}}} = r_{it} \qquad (11:17)$$

Proof

(1) $$r_{1t} = \frac{\sum Z_1 (Z_1 + Z_2 + \ldots Z_n)}{N(1)\sqrt{\frac{(Z_1 + Z_2 + \ldots Z_n)}{N}}}$$

(2) $$= \frac{\sum Z_1^2 + \sum Z_1 Z_2 + \ldots \sum Z_1 Z_n}{N\sqrt{\frac{\sum Z_1^2}{N}} + \frac{\sum Z_2^2}{N} + \frac{\sum Z_n^2}{N} + 2\frac{\sum Z_1 Z_2}{N} + 2\frac{\sum Z_1 Z_3}{N} + \ldots 2\frac{\sum Z_{(n-1)} Z_n}{N}}$$

(3) As $\frac{\sum Z^2}{N} = \sigma_z^2 = 1$, and $\frac{\sum Z_1 Z_j}{N} = r_{1j}$ (2) becomes:

$$\frac{1 + r_{12} + \ldots r_{1n}}{\sqrt{n + n(n-1)\overline{r_{ij}}}} = \frac{1 + (n-1)\overline{r_{1j}}}{\sqrt{n + n(n-1)\overline{r_{ij}}}}$$

(4) Squaring and recalling that $\overline{r_{1j}} = \overline{r_{ij}}$ (from (11:16)),

$$r_{1t}^2 = \frac{1 + (n-1)^2 r_{ij}^2 + 2(n-1)\overline{r_{ij}}}{n + n(n-1)\overline{r_{ij}}}$$

(5) Dividing numerator and denominator by n^2 gives:

$$r_{1t}^2 = \frac{\frac{1}{n^2} + \frac{(n-1)^2}{n^2}\overline{r_{ij}^2} + \frac{2(n-1)}{n^2}\overline{r_{ij}}}{\frac{1}{n} + \frac{n(n-1)}{n^2}\overline{r_{ij}}}$$

(6) As n = the number of items in the domain, n is infinitely large:

(a) Values divided by n become infinitely small and

(b) $\frac{(n-1)}{n}$ is virtually equal to 1. (5) thus becomes:

$$r_{1t}^2 = \frac{\overline{r_{ij}^2}}{\overline{r_{ij}}} = \overline{r_{ij}}$$

(7) Therefore:

$$r_{1t} = \sqrt{\overline{r_{ij}}}$$

The proportion of variance in an item accounted for by the correlation of an item with the true score will be $r_{1t}^2 = \overline{r_{ij}}$, as the total proportion of variance of the item will be 1.0, the proportion of true variance equals $\overline{r_{ij}}$, thus the reliability coefficient of an item equals $\overline{r_{ij}}$. Symbolising the reliability of an item as r_{ii}:

$$r_{ii} = \overline{r_{ij}} \qquad (11:18)$$

11. *The Correlation of a k-item Test with the True Score*

The correlation between a test made up of *k* items and the true score can be shown to be:

$$r_{kt} = \frac{k\bar{r}_{it}}{\sqrt{k + k(k-1)\overline{r_{ij}}}} \qquad (11:19)$$

Proof

(1) $$r_{kt} = \frac{\sum Z_1(Z_1 + Z_2 + \ldots Z_k)}{N(1)\sqrt{\frac{\sum(Z_1 + Z_2 + \ldots Z_k)^2}{N}}}$$

(2) $$= \frac{r_{t1} + r_{t2} + \ldots r_{tk}}{\sqrt{k + k(k-1)\overline{r_{ij}}}}$$

(It has been assumed that the average intercorrelation in *k* x *k* matrix will be the same as that in the *n* x *n* matrix. If *k* is reasonably large, and the items are a random sample from the domain this is a reasonable assumption.)

(3) So:

$$r_{kt} = \frac{k\overline{r_{it}}}{\sqrt{k + k(k-1)\overline{r_{ij}}}}$$

It also follows that:

(4) Squaring and recalling that $r_{it} = \sqrt{\overline{r_{ij}}}$ gives:

$$r_{kt}^{\ 2} = \frac{k^2\overline{r_{ij}}}{k + k(k-1)\overline{r_{ij}}}$$

(5) Dividing numerator and denominator by *k* gives:

$$r_{kt}^{\ 2} = \frac{k\overline{r_{ij}}}{1 + (k-1)\overline{r_{ij}}}$$

This last value on the left is the squared correlation between a *k* item test and true scores and is the reliability coefficient of a *k* item test, which, as in the section on parallel tests, can be symbolised r_{xx}, so:

$$r_{xx} = \frac{k\overline{r_{ij}}}{1+(k-1)\overline{r_{ij}}} = \frac{kr_{ii}}{1+(k-1)r_{ii}} \qquad (11:20)$$

The proportion of variance accounted for by the correlation with true scores will equal r_{xx}, and the residual error variance will be $1 - r_{xx}$, and the standard error of measurement will be, as before, $\sigma_x\sqrt{1-r_{xx}}$.

At this stage it should be noted that all of the above conclusions would have followed if instead of items, different sets of items randomly sampled from the domain had been used. These random sets would be the domain sampling equivalent of parallel tests, and are called randomly parallel tests, or randomly parallel composites. In the proofs, *Z* scores on tests would need to be substituted for *Z* scores on items.

Looking again at (11:20) which gives the reliability of a *k* item test, it will be seen that it is the item equivalent of (11:15). It is the Spearman-Brown formula.

Problems

A. Given a test consisting of *k* items with a reliability of .80, what would its reliability be if it was:

(1) doubled in length?
(2) lengthened tenfold?

B. (1) What would be the reliability of a test made up of all seven items with an average intercorrelation between items of .20?
(2) If the mean intercorrelation between items was .30, what would be the reliability of a 10 item test?
(3) If there were 50 items whose mean intercorrelation was .10, what would the reliability coefficient be?

Answers

A. (1) The Spearman-Brown formula states:

$$r_{cxcy} = \frac{r_{xx}}{1+(k-1)r_{xx}}$$

Substituting 2 for *k* and .80 for r_{xx} gives:

$$r_{cxcy} = \frac{2\times .80}{1+.80} = .89$$

(2) Substituting 10 for *k* and .80 for r_{xx} gives:

$$r_{cxcy} = \frac{10\times .80}{1+(9\times .80)} = .98$$

B. (1) Substituting .20 for $\overline{r_{ij}}$ and 7 for *k* in (11:19) gives:

$$r_{xx} = \frac{7\times .20}{1+(6\times .20)} = .64$$

(2) Substituting .30 for $\overline{r_{ij}}$ and 10 for *k* gives:

$$\frac{10\times .30}{1+(9\times .30)} = .81$$

(3) Substituting appropriately gives:

$$\frac{50\times .10}{1+(49\times .10)} = .85$$

12. *Coefficient Alpha and the Kuder-Richardson 20 Formula*

In step (4) of the proof of (11:19) the following formula occurred.

$$r_{xx} = \frac{k^2\overline{r_{ij}}}{k + k(k-1)\overline{r_{ij}}} \qquad (11:21)$$

The denominator of the right hand term is the *Z* score equivalent of the variance of a composite and can be turned into a raw score equivalent which will give:

$$r_{xx} = \frac{k^2\overline{r_{ij}\sigma_i\sigma_j}}{k\overline{\sigma}_i^2 + k(k-1)\overline{r_{ij}\sigma_i\sigma_j}} \qquad (11:22)$$

Where $\overline{r_{ij}\sigma_i\sigma_j}$ is the mean of the item covariances.
$\overline{\sigma}_i^2$ is the mean of the item variances.

The variance of the composite is of course the variance of the test so:

$$\sigma_x^2 = k\overline{\sigma}_i^2 + k(k-1)\overline{r_{ij}\sigma_i\sigma_j} \qquad (11:23)$$

It can be shown that:

$$\frac{\sigma_x^2 - k\overline{\sigma}_i^2}{k(k-1)} = \overline{r_{ij}\sigma_i\sigma_j} \qquad (11:24)$$

Proof

(1) $\sigma_x^2 = k\overline{\sigma}_i^2 + k(k-1)\overline{r_{ij}\sigma_i\sigma_j}$

(2) Subtracting $k\sigma_i^2$ gives:

$\sigma_x^2 - k\overline{\sigma}_i^2 = k(k-1)\overline{r_{ij}\sigma_i\sigma_j}$

(3) Dividing by $k(k-1)$ gives:

$$\frac{\sigma_x^2 - k\overline{\sigma}_i^2}{k(k-1)} = \overline{r_{ij}\sigma_i\sigma_j}$$

Returning to (11:22) and substituting:

(a) σ_x^2 for the denominator, and

(b) $\frac{\sigma_x^2 - k\overline{\sigma}_i^2}{k(k-1)}$ for $\overline{r_{ij}\sigma_i\sigma_j}$ in the numerator

It can be shown that:

$$r_{xx} = \frac{k}{k-1}\left(1 - \frac{\sum\sigma_i^2}{\sigma_x^2}\right) \qquad (11:25)$$

Proof

(1) $r_{xx} = \frac{k^2\left[\left(\sigma_x^2 - k\overline{\sigma}_i^2\right)/k(k-1)\right]}{\sigma_x^2}$

(2) $$r_{xx} = \frac{k^2}{k(k-1)}\left(\frac{\sigma_x^2 - k\overline{\sigma}_i^2}{\sigma_x^2}\right)$$

(3) In a *k* item test $\overline{\sigma}_i^2 = \frac{\sum\sigma_i^2}{k}$, therefore $k\overline{\sigma}_i^2 = \sum\sigma_i^2$, substituting accordingly and dividing by *k* gives:

$$r_{xx} = \frac{k}{k-1}\left(\frac{\sigma_x^2 - \sum \sigma_i^2}{\sigma_x^2}\right)$$

$$= \frac{k}{k-1}\left(1 - \frac{\sum \sigma_i^2}{\sigma_x^2}\right)$$

(11:25) is known as coefficient alpha and it relates to reliability of a test to the item variances.

As the ratio of item variances to total variance increases so reliability decreases. The ratio will equal:

$$\frac{\sum \text{item variances}}{\sum \text{item variances} \ + \ \sum \text{item covariances}}$$

As the correlation between items increases so the denominator becomes larger, because the item covariances increase, and the reliability of the test increases. This emphasizes yet again the importance of item intercorrelations in the domain sampling model of reliability.

When coefficient alpha is used with dichotomous items it becomes the Kuder-Richardson 20 formula. Recalling that the variance of a dichotomous item is *pq*, (11:25) , for dichotomous items, becomes:

$$r_{xx} = \frac{k}{k-1}\left(1 - \frac{\sum p_i q_i}{\sigma_x^2}\right) \qquad (11:26)$$

This is the method suggested by the domain sampling model for assessing the reliability of a test made up of dichotomous items.

13. *The Estimation of the Reliability Coefficient*

In basic text books of psychology three main ways of assessing reliability are mentioned. These are:

Parallel forms
Test-retest
Split-half

Usually Kuder-Richardson 20 is mentioned as well as a measure of internal consistency.

In the parallel form method of assessing reliability, two supposedly parallel forms of a test are administered to the same group of subjects with usually a week or two in between administrations. The correlation coefficient between tests is then computed and taken as an estimate of r_{xx}.

In the test-retest method, the same test is administered on two occasions to the same group of subjects, and the correlation between scores on the two occasions taken as an estimate of r_{xx}.

The split-half method involves splitting the test into two halves. Often the split is in terms of odd numbered items in one half and even numbered in the other. The data is usually gathered in one session, i.e. the whole test is administered. The two halves are then correlated, and the coefficient r_{xx} estimated/ As reliability depends on the number of measurements a correction has to be incorporated to allow for the halving of test length in computing the correlation. This is done by applying the Spearman-Brown formula with k = 2, as the full test is twice as long as each half.

All of these are attempts to form parallel tests, and in terms of true scores and error scores involve the following reasoning.

$$r_{xy} = \frac{\sum (t + e_1)(t + e_2)}{N\sigma_x\sigma_y}$$

$$= \left(\frac{\sum t^2}{N} + \frac{\sum te_1}{N} + \frac{\sum te_2}{N} + \frac{\sum e_1 e_2}{N} \right) \frac{1}{\sigma_x \sigma_y}$$

Because $\frac{\sum te_1}{N}$ and $\frac{\sum te_2}{N}$ and $\frac{\sum e_1 e_2}{N}$ are covariances of error terms with other variables and because by assumption correlation between error scores and other variables is zero, all of the terms equal zero.

So we obtain:

$$r_{xy} = \frac{\sum t^2 / N}{\sigma_x \sigma_y} = \frac{\sigma_t^2}{\sigma_x \sigma_y} = \frac{\sigma_t^2}{\sigma_x^2}$$

(If the tests are parallel forms $\sigma_x = \sigma_y$ so $\sigma_x \sigma_y = \sigma_x^2$). r_{xy} therefore equals the ratio of true variance to total variance, which equals r_{xx}. (If in doubt consult (11:5) and (11:6).)

Problems arise however when we consider what counts as error. Each method controls for certain sorts of error and not for others. A classification of sources of error is given below with brief examples. Most of these sorts of error arise out of interaction between subject and test, subject and situation, subject and tests, and so on, but they are listed as separate sources. In the following list 'transient' should be taken as meaning varying from day-to-day.

(1) Test errors - e.g. faults in standardisation, ambiguities in questions.

(2) Tester errors:
 (a) durable - e.g. constant errors of administration, permanent behavioural characteristics which affect the subject.
 (b) transient - careless errors of administration, transient states of behaviour which affect subject.

(3) Situational errors:
 (a) durable - permanent poor lighting, temperature, etc.

(b) transient - temporary lighting, temperature defects.

(4) Subject errors:
(a) durable - sensory defects, inappropriate cultural background for test, etc.
(b) transient - headaches, depression, elation, etc.
(c) fatigue.
(d) memory.
(e) practice effects.

In general the effects of durable errors can be detected by changing the test, the tester, and the situation. Test-retest correlations using as they do the same test, often the same tester and the same situation will be higher than they would if these factors were all changed.

Transient errors on the other hand can be detected by changing the occasions. Correlations between tests given on different occasions will be lowered by day-to-day fluctuations, while correlations between tests given on the same day will be higher. The danger of changing occasions is of course, that in the time interval between occasions true change can occur, which will be treated as error. The relationship of these factors to the different methods of measuring reliability is shown below in Table 11:1.

TABLE 11:1 ERRORS IN DIFFERENT METHODS OF RELIABILITY ASSESSMENT

(1) Parallel forms (same day):
No detection of day-to-day fluctuations. Differential proneness to fatigue. Practice effects probably more pronounced than after a time interval. Durable errors not detected, except for test errors.

(2) Parallel forms (different occasions):
Real change can be confused with error. Durable errors not detected except for test errors. Practice effects.

(3) Test-retest (same day):
No detection of day-to-day fluctuations. Differential proneness to fatigue. Memory produced errors. Strong practice effects. No durable errors detected.

(4) Test-retest (different occasions):
Real change can be confused with error. Durable errors not detected. Memory produced errors. Practice effect.

(5) Split-half
No detection of day-to-day fluctuations. Durable errors not detected. Different splits will give different results.

N.B. In all cases there is the question of how good an approximation to parallel tests has been obtained.

Some cautions must of course go with the table. For example, fatigue effects will operate within a single test, but they are presumably an increasingly important factor as the test gets longer. But the table does give an idea of the errors detected by different methods of reliability assessment. The choice of method will depend on which type of error is considered most important to detect.

The Kuder-Richardson 20 formula uses one test and one testing session. It will be recalled that we derived it from a formula using inter-item correlations, and further that the average inter-item correlations, and further that the average inter-item correlation was the value of interest. While day-to-day fluctuations will affect total score on tests, it is not so likely that they will affect inter-item correlations. If the test is of any length at all the effect of the fluctuations might be expected to cancel one another out. Similarly factors such as the effects of ambiguities in items are likely to be minimized because of the large number of intercorrelations taken into account by the average value. For these reasons it might be argued that the Kuder-Richardson 20 formula is less susceptible than the others to the effects of changes of occasion, situation, etc. In any case in terms of domain sampling theory K-R 20's use of inter-item correlations makes it the best estimate of the correlation between scores on the test and true scores.

However, the domain sampling theory assumes that the sample of items in the test is a random or at least a representative sample, and it is obvious from consideration of the way tests are made that:

(1) test items are invented by the test maker,
(2) they are not a random sample from the domain, and
(3) that therefore they may or may not be representative.

Further, if the test is composed of heterogenous items K-R 20 will yield a low value. For example if the test was designed to predict sales ability, and was composed of personality and intelligence test items, we know that personality variables and intelligence tend to be uncorrelated, so the items would have low intercorrelations and thus K-R 20 would be low. The other estimates of reliability could still be high in this sort of situation, but K-R 20 would be much affected by lack of homogeneity amongst items. Digressing slightly, it can be argued that conglomerate tests are undesirable. They may be useful for purely empirical and predictive purposes but they are of no scientific value. In the situation outlined we should use separate tests of intelligence and personality variables. K-R 20 could then be quite successfully applied to each of the separate tests.

14. *Standard Error of Measurement*

It has been shown that in both the true and error score model, and the domain sampling model that the variance of error scores is equal to:

and that the standard error of measurement is:

$$\sigma_{meas} = \sigma_x \sqrt{1 - r_{xx}} \qquad (11:27)$$

If it can be assumed that the value of σ_{meas} is the same for all score levels, it can be used to estimate the range within which the true score is likely to fall. Consider the distribution below:

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

Figure 11.1 The distribution of obtained scores about the true score

This is the distribution of an individual's obtained scores about his true score. The standard deviation of this distribution is the standard error of measurement. As the distribution is assumed to be a normal one, 68.26 per cent of obtained scores will lie within the range + 1 to - 1 σ_{meas}; 95.44 per cent within the range $+2$ to $-2\sigma_{meas}$; and 99.73 per cent within the range $+3$ to $-3\sigma_{meas}$.

Given an obtained score, which must come from somewhere in the distribution, it is possible to work out the range within which the true score lies. As error can be positive or negative we do not know whether the obtained score is above or below the true score, so in our estimate we will have to allow for the possibility that it is above and the possibility that it is below.

Let us suppose that we want to be 95 per cent certain of including the true score in a range that we specify. Then we would need to cover the range indicated in the distribution below in Figure 11.1.

We know from consulting tables for the normal curve that 95 per cent of cases be within the range + 1.96 to -1.96σ, so in the case of obtained scores 95 per cent will be within the range +1.96 σ_{meas} to $-1.96\sigma_{meas}$.

To allow for the possibility that our obtained score is above the true score we subtract from the obtained score 1.96 σ_{meas}. Only 2.5 per cent of obtained scores above the true score will be further away from it than this. Similarly to allow for the possibility that the obtained score is below the true score we must add $1.96\sigma_{meas}$ to it. Only 2.5 per cent of obtained scores below the true score would be further away from it than this So by finding the range: obtained score $\pm 1.96\sigma_{meas}$, we can state that it is 95 per cent certain that the true score lies within that range.

A step by step analysis might be useful here. To find the range within which the true score is likely to lie:

Step 1. Calculate σ_{meas} which equals $\sigma\sqrt{1-r_{xx}}$.

Step 2. Decide on how confident you want to be of including the true score within the specified range.

Step 3. Work out the upper point of the range by adding the appropriate number of σ_{meas} to the obtained score.

Step 4. Work out the lower point by subtracting the appropriate number of σ_{meas} from the obtained score.

For 95 per cent confidence	$\pm 1.96\sigma_{meas}$
For 99 per cent confidence	$\pm 2.58_{meas}$
For 99.9 per cent confidence	$\pm 3.29_{meas}$

The two assumptions made on these calculations are:

(1) that σ_{meas} is the same for every individual;
(2) that the distribution of obtained scores is normal.

It is possible that these assumptions are not always justified. In some models $\sigma\sqrt{1-r_{xx}}$ is the value of the average σ_{meas}, and in some models the distribution of errors is skewed towards the mean score for all individuals. However in most circumstances the use of σ_{meas} in the fashion outlined above will be a reasonable approximation.

Problems

A. If a student obtains a score of 120 on test with a mean of 100; a standard deviation of 12; and a reliability coefficient of .89, what is the range within which, with 95 per cent confidence, his true score lies?

B. A patient is given a test of anxiety with a reliability coefficient of .64. He obtains a score at the 75th percentile, which surprises the psychologist, who thought that the patient's anxiety would be at the 97.5th percentile. What is the probability of someone with a test score at the 75th percentile having a true score at the 97.5th?

Answers

A. (1) $\sigma_{meas} = \sqrt{1-.891=12}\sqrt{.11} = 4.$

(2) 95 per cent confidence limits $= \pm 1.96\sigma_{meas} = 112.16 - 127.84$

B. (1) 97.5th percentile $= Z, +1.96.$

(2) σ_{meas} in Z score terms = $\sqrt{1-.64} = .60.$

(3) 75th percentile = $Z, +.67$, so the obtained score is $2.15\sigma_{meas}$ away from a true score at the 97.5th percentile.

(4) Less than two people in a hundred with a true score at the 97.5th percentile would be expected to have an obtained score at the 75th percentile.

15. *Obtaining a Test of a given Reliability*

It will be recalled that according to the Spearman-Brown formula (11:15), (11:19), the effects of lengthening a test can be estimated from the formula:

$$r_{x'x'} = \frac{kr_{xx}}{1+(k-1)r_{xx}}$$

Where $r_{x'x'}$ is the reliability of the lengthened test; *k* is the number of items in the new test divided by the number of items in the original test; and r_{xx} is the reliability of the original test.

To answer the question of how long a test needs to be to attain a given reliability it is necessary to manipulate the Spearman-Brown formula to give:

$$k = \frac{r_{x'x'}(1-r_{xx})}{r_{xx}(1-r_{x'x'})} \qquad (11:28)$$

Proof

(1) $r_{x'x'} = \frac{kr_{xx}}{1+(k-1)r_{xx}}$

(2) Multiplying both sides by $1+(k-1)r_{xx}$ gives:

$$r_{x'x'}\left[1+(k-1)r_{xx}\right] = kr_{xx}$$

(3) $= r_{x'x'} + r_{x'x'}kr_{xx} - r_{x'x'}r_{xx} = kr_{xx}$

(4) Subtracting $r_{x'x'}kr_{xx}$ from both sides gives:

$$r_{x'x'} - r_{x'x'}r_{xx} = kr_{xx} - r_{x'x'}kr_{xx} =$$
$$r_{x'x'}(1-r_{xx}) = k(r_{xx} - r_{x'x'}r_{xx})$$

(5) Dividing both sides by $r_{xx} - r_{x'x'}r_{xx}$ gives:

To use this formula let us suppose that we have a test of anxiety, consisting of 20 items, which has a reliability coefficient of .50, and that a test with a reliability of .90 is required. How many items will be needed for the more reliable test?

Substituting values in (11:26) gives:

$$k = \frac{.90(1 - .50)}{.50(1 - .90)} = \frac{.45}{.05} = 9$$

To achieve a test with a reliability of .90, it will be necessary to lengthen the existing one nine-fold. Thus 180 items will be needed to give a reliability coefficient of .90.

Problems

A. Given a test with a mean of 100, standard deviation 15, and r_{xx} .90, what is the value of σ_{meas}?

B. If someone obtains a score of 110 or above on such a test what is the probability of his true score being 100?

C. A questionnaire has a reliability coefficient of .70, and consists of 10 items. By how much will it need to be lengthened to make its reliability .80?

D. A test has a reliability coefficient of .90, but consists of 180 items. This is considered too long a test for most subjects, and it is proposed to reduce its length to 90 items. What will be its new reliability coefficient?

Answers

A. $\sigma_{meas} = \sigma\sqrt{1 - r_{xx}} = 15\sqrt{1 - .90} = 15 \times .33 = 5.$

B. 110 is two σ_{meas} above 100. The probability of someone with a true score of 100 obtaining a score of 110 is .023.

C. Substituting values in (11:26) gives:

$$k = \frac{.80(1 - .70)}{.70(1 - .80)} = \frac{.24}{.14} = 1.71$$

The new test will need to be 1.71 times as long.

D. For this problem (11:15) or (11:19) should be used. Substituting gives:

$$r_{x'x'} = \frac{.50 \times .90}{1 - (.50 \times .90)} = \frac{.45}{.55} = .82$$

Validity

1. *Types of Validity*

Traditionally four types of validity have been described:

(1) Face validity

(2) Content validity

(3) Empirical validity
  (a) concurrent
  (b) predictive

(4) Construct validity

Face validity is concerned with whether a test looks as if it measures what it is supposed to measure. Of course, tests may look as though they measure something and not really measure it at all, so face validity is of little use except from the consumer relations point of view. Suppose that there were two tests which were equally good at discriminating between neurotics and normals, and that one test was made up of questions about artistic and literary preferences, and the other of questions about nightmares, anxieties, and the like. The second test has obvious face validity and for this reason might be more acceptable to patients than the first. So, other things being equal, face validity can be a deciding factor in which of a number of test should be used.

Content validity is the extent to which a test adequately samples the universe or domain of items which it is supposed to measure. It is of importance in the field of achievement testing. If, for example, it is desirable to know whether a given 12 year old is above or below average in arithmetical attainment, it is necessary to have in the test an adequate sample of the types of arithmetic operations and problems that average 12 year olds can deal with. To the extent to which the test was an adequate sample, to that extent it would have content validity.

Empirical validity is determined by directly relating test scores or other predictors to the criterion of interest. In the case of concurrent validity the relationship between a test and a currently available criterion is assessed, while in the case of predictive validity the criterion does not become available until a later date. Empirical validity is usually expressed in the form of a correlation coefficient, but sometimes a test of significance between the scores of criterion groups is reported instead. The danger with the latter method is that significant differences can in fact be found even when the degree of association between predictor and criterion is low. Fortunately methods of working out an index of association from tests of significance are available. Some of these will be discussed in Section 5 of this chapter.

Construct validity is assessed by seeing whether scores on a test which purports to measure a given trait, are associated with behavioural differences which a theory says should be associated with the trait. One of the best examples of this was the use of the Taylor Manifest Anxiety Scale by Spence, Taylor and their associates. In these researches it was hypothesised that anxiety was a drive state. In terms of Hull-Spence theory certain predictions could be made about the relationship between strength of drive and performance. If the Taylor Manifest Anxiety Scale measured drive then high scorers on the scale should differ from low scorers on various performance indices. Much research was generated by the theory and the record of success in predicting performance in situations where there were no competing responses was good, thus suggesting some degree of construct validity for the Taylor MAS. Another good example of construct validity is the work of McClelland and his associates on Need Achievement.

These brief descriptions will have to serve for all types of validity, except empirical validity. With regard to this some factors affecting the size of the correlation between two variables will be considered, as will the derivation of measures of association from some tests of significance.

2. *The Effects of Unreliability on the Correlations between Variables*

In terms of true and error scores it is fairly easy to work out the effects of unreliability on the correlation between two variables. If the interest is in the correlation between two variables *X* and *Y* with the effects of unreliability cancelled out from both *X* and *Y*, then the correlation between true scores on *X* and true scores on *Y* is required, and this will be given by the formula:

$$r_{x_t y_t} = \frac{\sum(x - e_x)(y - e_y)}{N\sigma_{x_t}\sigma_{y_t}} \qquad (12{:}1)$$

where: $r_{x_t y_t}$ is the correlation between true scores on *X* and *Y*,
σ_{x_t} is the standard deviation of true scores on *X*,
σ_{y_t} is the standard deviation of true scores on *Y*.

The standard deviations of true scores are derived from:

$$\sigma_{x_t} = \sigma_x\sqrt{r_{xx}} \qquad (12{:}2)$$

Proof

(1) $$r_{xx} = \frac{\sigma^2_{x_t}}{\sigma^2_x}$$

(2) Multiplying both sides by σ^2_x gives:

$$\sigma^2_x r_{xx} = \sigma^2_{x_t}$$

(3) Taking the square root of both sides gives:

$$\sigma_x\sqrt{r_{xx}} = \sigma_{x_t}$$

Returning to (12:1) and inserting the appropriate formulae for σ_{x_t} and σ_{y_t} gives:

$$r_{x_ty_t} = \frac{r_{xy}}{\sqrt{r_{xx}}\sqrt{r_{yy}}} \qquad (12:3)$$

Proof

(1)

$$r_{x_ty_t} = \frac{\sum(x-e_x)(y-e_y)}{N\sigma_x\sqrt{r_{xx}}.\sigma_y\sqrt{r_{yy}}}$$

(2) Dividing numerator and denominator by *N* gives:

$$\frac{\left(\sum xy + \sum e_xe_y - \sum xe_y - \sum ye_x\right)/N}{\sigma_x\sigma_y\sqrt{r_{xx}}\sqrt{r_{yy}}}$$

(3) The numerator now consists of a number of covariance terms so (2) can be written:

$$\frac{r_{xy}\sigma_x\sigma_y + r_{e_xe_y}\sigma_{e_x}\sigma_{e_y} - r_{xe_y}\sigma_x\sigma_{ey} - r_{ye_x}\sigma_y\sigma_{e_x}}{\sigma_x\sigma_y\sqrt{r_{xx}}\sqrt{r_{yy}}}$$

(4) By assumption the correlation between error scores and other scores is zero, therefore (3) becomes:

(5) Dividing numerator and denominator by $\sigma_x\sigma_y$ gives:

$$r_{x_ty_t} = \frac{r_{xy}}{\sqrt{r_{xx}}\sqrt{r_{yy}}}$$

the use of this formula is known as correcting for attenuation, and it gives the correlation between a p0erfectly reliable measure of *X* and a perfectly reliable measure of *Y*.

While this correction procedure is of value for theoretical purposes, its main usefulness in terms of empirical validity is that it gives the upper limit of the correlation obtainable between two variables. Its computation can therefore help decide whether it is worth taking steps to increase the reliability of predictor or criterion.

In real life situations, e.g. predicting essay type exam marks, or predicting occupational success, it is often difficult to do much about the criterion as this is frequently beyond the investigator's control;. This raises the problem of how well a perfectly reliable predictor would correlate with an unreliable criterion. Again the solution is fairly simple:

$$r_{x_t y} = \frac{r_{xy}}{\sqrt{r_{xx}}} \tag{12:4}$$

Proof

(1) $$r_{x_t y} = \frac{\sum(x - e_x)y}{N\sigma_x\sqrt{r_{xx}}\sigma_y}$$

(2) $$= \frac{\left(\sum xy - \sum e_x y\right)/N}{\sigma_x\sigma_y\sqrt{r_{xx}}}$$

(3) The numerator consists of two covariance terms one of which equals zero:

$$\frac{r_{xy}\sigma_x\sigma_y - r_{e_x y}\sigma_{e_x}\sigma_y}{\sigma_x\sigma_y\sqrt{r_{xx}}} = \frac{r_{xy}\sigma_x\sigma_y}{\sigma_x\sigma_y\sqrt{r_{xx}}}$$

(4) Dividing numerator and denominator by $\sigma_x\sigma_y$ gives:

$$r_{x_t y} = \frac{r_{xy}}{\sqrt{r_{xx}}}$$

This formula gives the relationship between a perfectly reliable variable and a variable still contaminated by error. The correction for error has only been carried out on one variable.

Problems

A. If the reliability coefficient of ratings of improvement in anxiety is .36, and a measure of skin potential before treatment which has a reliability of .64, correlates .24 with ratings of improvement, is it worth trying to make the procedures more reliable in an attempt to use skin potential to predict response to treatment? Work out the correlation which would be obtained with two error-free measures.

B. An intelligence test with a reliability coefficient of .81 correlates .45 with exam marks. If the test was made perfectly reliable what would the correlation with exam marks be?

Answers

A. Formula (12:3) is needed here. Substituting the appropriate value gives:

$$r_{x_t y_t} = \frac{.24}{\sqrt{.36}\sqrt{.64}} = .50$$

(The correlation would be raised from .24 to .50 which is a sizeable increase. Problems of how long the test would need to be, and how long the rating scale would need to be must now decide the issue. These could be solved using Kuder-Richardson 20. The other factors affecting the decision would be availability of other predictors, and the incremental validity of skin potential, i.e. the extent to which it added to predictions made from other sources.)

B. Solution of this problem requires formula (12:4), inserting appropriate values gives:

$$r_{x_t y} = \frac{.45}{\sqrt{.81}} = .50$$

3. *The Effects of Unreliability on the Highest Correlation Possible between Two Variables*

This topic has in fact been dealt with implicitly in the previous section, but because of the common belief that a test cannot have a correlation with another test, which is higher than its reliability coefficient, this special section has been inserted.

It will be recalled that the variance of a test can be split into two components.

$$\sigma_x^2 = \sigma_t^2 + \sigma_e^2$$

Error scores do not correlate with other scores, so the proportion of error variance accounted for by other variables must be zero. This leaves true variance (σ_t^2) which could be accounted for by another variable. The highest coefficient of determination possible is therefore:

$$\frac{\sigma_t^2}{\sigma_x^2}$$

The square root of the highest coefficient of determination will be the highest correlation which that test can have with another:

$$r_{xy}\ (\max) = \sqrt{\frac{\sigma_t^2}{\sigma_x^2}} = \sqrt{r_{xx}} \qquad (12:5)$$

So the highest correlation an unreliable test can have with another variable is equal to the index of reliability, which is the square root of the reliability coefficient.

4. *The Effects of Restricted or Increased Range on the Magnitude of Correlation Coefficients*

In Chapter 5 this topic was touched on in diagrammatic form. Now a more formal approach will be attempted before considering the effects of change in range it is necessary to make two assumptions:

(a) that the standard error of estimate is the same throughout the total range, and

(b) that the slope of the regression line is constant throughout the whole range.

It will be recalled that the standard error of estimate was symbolised as $\sigma_{y.x} = \sigma_y \sqrt{1 - r_{xy}^2}$.

In the following discussion there will be the available data whose standard error of estimate will be symbolised as $\sigma_{y.x}$, and this will be contrasted with the standard error of estimate for the changed range $\sigma'_{y.x}$. Similarly σ_x, σ_y will refer to available data and σ'_y, σ'_y to standard deviations of the changed range. It will now be shown that:

$$\sigma'_y = \frac{r_{xy}\sigma_y\sigma'_x}{r'_{xy}\sigma_x} \qquad (12:6)$$

Proof

(1) By assumption (b) above:

$$b_{y.x} = r_{xy}\frac{\sigma_y}{\sigma_x} = r'_{xy}\frac{\sigma'_y}{\sigma'_y}$$

(2) Dividing (1) by r'_{xy} gives:

$$\frac{r_{xy}\sigma_y}{r'_{xy}\sigma_x} = \frac{\sigma'_y}{\sigma'_x}$$

(3) Multiplying by σ'_x gives:

$$\sigma'_y = \frac{r_{xy}\sigma_y\sigma'_x}{r'_{xy}\sigma_x}$$

Further, by assumption (a) above, the standard error of estimate is constant throughout the range so:

$$\sigma'_{y.x} = \sigma_y\sqrt{1-r_{xy}^2} = \sigma'_y\sqrt{1-r'^2_{xy}} = \frac{r_{xy}\sigma_y\sigma'_x}{r'_{xy}\sigma_x}\sqrt{1-r'^2_{xy}} \qquad (12{:}7)$$

(This last term is obtained by substituting from (12:6).) Given these formulae it can be shown that:

$$r'_{xy} = \frac{r_{xy}\left(\dfrac{\sigma'_x}{\sigma_x}\right)}{\sqrt{1-r_{xy}^2+r_{xy}^2\left(\dfrac{\sigma'^2_x}{\sigma_x^2}\right)}} \qquad (12{:}8)$$

Proof

(1) From (12:7):

$$\sigma_y\sqrt{1-r_{xy}^2} = \frac{r_{xy}\sigma_y\sigma'_x}{r'_{xy}\sigma_x}\sqrt{1-r'^2_{xy}}$$

(2) Squaring both sides gives:

$$\sigma_y^2\left(1-r_{xy}^2\right) = \frac{r_{xy}^2\sigma_y^2\sigma'^2_x}{r'^2_{xy}\sigma_x^2}\left(1-r'^2_{xy}\right)$$

(3) Dividing by σ_y^2 gives:

$$1-r_{xy}^2 = \frac{r_{xy}^2\sigma'^2_x}{r'^2_{xy}\sigma_x^2}\left(1-r'^2_{xy}\right)$$

(4) Multiplying by $\dfrac{\sigma_x^2}{r_{xy}^2\sigma'^2_x}$ gives:

$$\frac{\sigma_x^2\left(1-2_{xy}^2\right)}{r_{xy}^2\sigma'^2_x} = \frac{1-r'^2_{xy}}{r'^2_{xy}}$$

(5) The right hand term equals:

$$\frac{1}{r'^2_{xy}} - \frac{r_{xy}^2}{r'^2_{xy}} = \frac{1}{r'^2_{xy}} - 1$$

(6) Adding 1 to both sides gives:

$$1+\frac{\sigma_x^2\left(1-r_{xy}^2\right)}{r_{xy}\sigma_x'^2}=\frac{1}{r_{xy}'^2}$$

(7) As a value divided by itself equals unity, $\dfrac{r_{xy}^2\sigma_x'^2}{r_{xy}^2\sigma_x'^2}=1$.

Substituting this on the left hand term of (6) gives:

(8) Inverting both sides gives:

$$\frac{r_{xy}^2\sigma_x'^2}{r_{xy}^2\sigma_x'^2+\sigma_x^2\left(1-r_{xy}^2\right)=r_{xy}'^2}=r_{xy}'^2$$

(9) Dividing the numerator and denominator of the left hand term by σ_x^2 gives:

$$\frac{r_{xy}^2\left(\dfrac{\sigma_x'^2}{\sigma_x^2}\right)}{r_{xy}^2\left(\dfrac{\sigma_x'^2}{\sigma_x^2}\right)=1-r_x^2}=r_{xy}'^2$$

(10) The square root will give:

$$r'_{xy} = \frac{r_{xy}\left(\dfrac{\sigma'_x}{\sigma_x}\right)}{\sqrt{1-r_{xy}^2+{}_{xy}^2\left(\dfrac{\sigma'^2_x}{\sigma_x^2}\right)}}$$

This formula gives the correlation between *X* and *Y* following a change in the range of *X*. Examination of the numerator shows that the correlation coefficient of the old range is multiplied by the ratio of the standard deviation of the new range to the standard deviation of the original range. Thus if the range is increased σ'_x will be greater than σ'_x will be greater than σ_x and r'_{xy} will be greater than r_{xy}. While if the range is decreased σ'_x will be less than σ_x, and r'_{xy} will be less than r_{xy}.

Problems

A. If *X* has a standard deviation of 10 and $r_{xy} = .50$, what will be the value of r'_{xy} if the range of *X* scores is so increased that they have a standard deviation of 20?

B. Given *X*, as above, with a standard deviation of 10 and $r_{xy} = .50$. What will the value of r'_{xy} be if the range of scores on *X* is restricted so that $\sigma'_x = 5$?

Answers

A.

$$r'_{xy} = \frac{r_{xy}\left(\frac{\sigma'_x}{\sigma_x}\right)}{\sqrt{1 - r_{xy}^2 + r_{xy}^2\left(\frac{\sigma'^2_x}{\sigma^2_x}\right)}} = \frac{.50\left(\frac{20}{10}\right)}{\sqrt{1 - .25 + .25\left(\frac{400}{100}\right)}} = \frac{1.00}{\sqrt{1.75}} = \frac{1.00}{1.32} = .76$$

B.

$$r'_{xy} = \frac{.50\left(\frac{5}{10}\right)}{\sqrt{1 - .25 + .25\left(\frac{25}{100}\right)}} = \frac{.25}{\sqrt{.81}} = \frac{.25}{.90} = .28$$

5. *Measures of Association from Tests of Significance*

It was mentioned in the introduction to this Chapter that it is possible to convert the results of tests for the significance of differences into measures of association. Some of these will be briefly described.

It will be recalled that the correlation coefficient is equal to the square root of the coefficient of determination, which is equal to the proportion of variance accounted for. Most significance tests can be converted into such a ratio. As a first step in the process, suppose that *j* sets of scores are available. The deviation score of the *i*th individual in the *j*th group will equal:

$$X_{ij} - M = \left(X_{ij} - \overline{X}_j\right) + \left(\overline{X}_j - M\right) \qquad (12:9)$$

Where X_{ij} is the score of the *i*th individual in the *j*th group.
$\overline{X}_j$ is the mean of scores in the *j*th group.
M is the mean of all scores.

From this it follows that the sum of squares can be split into two components, where $n_j =$ the number of cases in a group.

$$SS_x = SS_{(within)} + SS_{(between)} \qquad (12:10)$$

Or

$$\sum_{i=1}^{N}\left(X_{ij} - M\right)^2 = \sum_{j=1}^{J}\sum_{i=1}^{nj}\left(X_{ij} - \overline{X}_j\right)^2 + \sum_{j=1}^{J}\left(n_j\left[\overline{X}_j - M\right]^2\right)$$

Proof

(1) $X_{ij} - M = \left(X_{ij} - \overline{X}_j\right) + \left(\overline{X}_j - M\right)$

(2) $\left(X_{ij} - M\right)^2 = \left[\left(X_{ij} - \overline{X}_j\right) + \left(\overline{X}_j - M\right)\right]^2$

(3) $= \left(X_{ij} - \overline{X}_j\right)^2 + \left(\overline{X}_j - M\right)^2 + 2\left(X_{ij} - \overline{X}_j\right)\left(\overline{X}_j - M\right)$

(4) Summing across the individuals within a group and remembering that within a group $\left(\overline{X}_j - M\right)$ is a constant gives

$$\sum_{i=1}^{n_j}\left(X_{ij} - M\right)^2 = \sum_{i=1}^{n_j}\left(X_{ij} - \overline{X}_j\right)^2 + n_j\left(\overline{X}_j - M\right)^2 + 2\sum\left(X_{ij} - \overline{X}_j\right)\left(\overline{X}_j - M\right)$$

(5) As $\sum\left(X_{ij} - \overline{X}_j\right)$ is the sum of deviations of scores in a group from the mean of that group it equals zero, thus:

$$\sum_{i=1}^{n_j}\left(X_{ij} - M\right)^2 = \sum_{i=1}^{n_j}\left(X_{ij} - \overline{X}_j\right)^2 + n_j\left(\overline{X}_j - M\right)^2$$

(6) Summing across of *J* groups gives:

$$\sum_{i=1}^{N}\left(X_{ij} - M\right)^2 = \sum_{j=1}^{J}\sum_{i=1}^{n_j}\left(X_{ij} - \overline{X}_j\right)^2 + \sum_{j=1}^{J}\left[n_j\left(\overline{X}_j - M\right)^2\right]$$

These values are referred to as follows:

$\sum_{i=1}^{N}\left(X_{ij} - M\right)^2$ = Total Sum of Squares = $SS_{(total)}$

$\sum_{j=1}^{J}\sum_{i=1}^{n_j}\left(X_{ij} - \overline{X}_j\right)^2$ = Within Sum of Squares

$= SS_{(within)}$

$$\sum_{j=1}^{J}\left(n_j\left[\overline{X}_j - M\right]^2\right) = \text{Between Sum of Squares}$$
$$= SS_{(between)}$$

Dividing (12:10) by *N* gives:

$$\sigma_x^2 = \frac{SS_{(within)}}{N} + \frac{SS_{(between)}}{N} \qquad (12:11)$$

Thus the variance is split into two parts, one arising from differences within groups, i.e. deviations of scores from group means, and one part due to differences between groups, i.e. arising from differences between group means and the grand mean.

The proportion of variance due to differences between groups will be given by:

$$\frac{SS_{(between)}}{SS_{(total)}} \qquad (12:12)$$

This is the proportion of variance accounted for by differences between groups, and is analogous to a coefficient of determination. Its square root will therefore be analogous to r_{xy}, and is known as eta, symbolised η

$$\eta = \sqrt{\frac{SS_{(between)}}{SS_{(total)}}} \qquad (12:13)$$

The value of eta can readily be obtained from an analysis of variance table by dividing the between sum of squares by the total sum of squares and taking the square root. The value can also be found from the *F* ratio by means of the formula:

$$\eta = \sqrt{\frac{F\left(df_b\right)}{F\left(df_b\right)+\left(df_w\right)}} \qquad (12:14)$$

where

F is the *F* ration

(df_b) is the number of degrees of freedom between groups
(df_w) is the number of degrees of freedom within groups

Proof

(1) $\eta^2 = \frac{SS_{(between)}}{SS_{(total)}}$

(2) $\frac{SS_{(between)}}{SS_{(total)}} = \frac{SS_{(between)}}{SS_{(between)} + SS_{(within)}}$

(3) But $F = \frac{\text{between variance estimate}}{\text{within variance estimate}}$

$$= \frac{SS_{(between)}\big/\left(df_b\right)}{SS_{(within)}\big/\left(df_w\right)} = \frac{\sigma^2_{(between)}}{\sigma^2_{(within)}}$$

(4) From this formula it can be seen that:

(a) $\left(df_b\right)\sigma_b^2 = SS_{(between)}$

(b) $\left(df_w\right)\sigma_w^2 = SS_{(within)}$

(c) $\sigma_b^2 = \sigma_w^2 F$.

(5) Substituting these values in (2) gives:

$$\frac{SS_{(between)}}{SS_{(within)}} = \frac{\sigma_w^2 F(df_b)}{\sigma_w^2 F(df_b) + \sigma_w^2 (df_w)}$$

(6) Dividing (5) by σ_w^2 gives:

$$\frac{SS_{(between)}}{SS_{(total)}} = \frac{F(df_b)}{F(df_b + (df_w))}$$

(7) The square root of this is eta.

If only two groups had been involved it is more likely that a *t*-ratio would have been reported. As $t = \sqrt{F}$ the formula can be easily adapted so:

$$\eta = \sqrt{\frac{t^2}{t^2 + df}} \qquad (12{:}15)$$

where *t* is the *t* ratio computed as usual and *df* is the number of degrees of freedom for *t*, which equals *N* - 2, here *N* is the total number of observations.

Proof

(1) $\eta = \sqrt{\frac{F(df_b)}{F(df_b) + (df_w)}}$

(2) $F = t^2$ so

$$\eta = \sqrt{\frac{t^2(df_b)}{t^2(df_b) + (df_w)}}$$

(3) (df_b) = the number of groups minus 1, so in the case of a *t* test where the number of groups is 2, (df_b) = 1

(4) (df_w) = total $df - (df_b)$, and total $df = N - 1$, so (df_w) in the case of a *t* test = *N* - 1 - 1 = *N* - 2

(5) But *N* - 2 = number of degrees of freedom for a *t* test.

(6) Substituting from (3), (4), and (5) in (2) gives:

$$\eta = \sqrt{\frac{t^2(1)}{t^2(1) + df}} = \sqrt{\frac{t^2}{t^2 + df}}$$

Strictly speaking the value of eta so obtained will apply only to the sample, but a correction can be applied to estimate the population value. The correction consists of subtracting $(df_b)\sigma^2{}_{(within)}$ from the numerator of step (5) of the proof of (12:14). In step (6) everything was divided by $\sigma^2{}_{(within)}$, so the corrected value of eta, called epsilon, is:

$$\varepsilon = \sqrt{\frac{F(df_b) - (df_b)}{F(df_b) + (df_w)}} = \sqrt{\frac{(df_b)(F-1)}{F(df_b) + (df_w)}} \qquad (12:16)$$

In most validational and normative studies the value of (df_b) will be small (often only two groups are used), and the value of (df_w) will be high (usually hundreds of subjects are used), so the difference between eta and epsilon will be negligible.

Problems

A. The distribution of scores obtained from 6 diagnostic groups on a new test is subjected to analysis of variance with the following results.

Source	*SS*	*df*	*Variance Estimate*	*F*
Diagnosis (between)	100	5	20	10
Within groups	200	100	2	

Work out an index of association between test and diagnosis.

B. In a comparison of the scores of schizophrenics and neurotics on a new test a *t* ratio of 5.0 is found, with a sample of 302 neurotics and 300 schizophrenics this is significant at well beyond the .001 level. Would the test be very valuable for diagnostic purposes?

C. In a pilot study 20 employees are placed into four grades of success, and an analysis of variance of scores on a test administered previously at entry to employment is carried out with the following results.

Source	*SS*	*df*	*Mean Square*	*F*
Grades (between)	27	3	9	9.00
Within	16	16	1	

What are the values of η and ε for the above data?

Answers

A. $$\eta=\sqrt{\frac{F(df_b)}{F(df_b+df_w)}}=\sqrt{\frac{10\times 5}{(10\times 5)+100}}=\sqrt{.30}=.55$$

B. $$\eta=\sqrt{\frac{t^2}{t^2+df}}=\sqrt{\frac{25}{25+600}}=\sqrt{.04}=.20$$

C. $$\eta=\sqrt{\frac{9\times 3}{(9\times 3)+16}}=\sqrt{\frac{27}{43}}=\sqrt{.63}=.79$$

$$\varepsilon=\sqrt{\frac{(df_b)(F-1)}{F(df_b)+(df_w)}}=\sqrt{\frac{3\times 8}{(3\times 9)+16}}$$

$$=\sqrt{\frac{24}{43}}=\sqrt{.56}=.75$$

6. *Another Look at Reliability*

In the previous section it has been demonstrated that the proportion of variance accounted for by differences between groups can be assessed from an analysis of variance, and transformed into a measure of association.

Suppose that instead of a group of scores consisting of scores from different individuals in a given condition, the group of scores consisted of a number of scores on a given test for one individual. Each group would contain *m* scores from the one individual and there would be *n* individuals. Differences between the *n* groups would now be differences between individuals. In terms of true and error score theory the group mean should now approximate the true score of an individual, and differences between groups should equal differences between true scores and thus be true variance $-\ \sigma_t^2$.

If the test was perfectly reliable then the individual would get the same score on every testing and all scores within a group would be the same. In so far as scores within a group differ they represent error variance $-\ \sigma_e^2$. Thus $SS_{(between)}$ would be the sum of squares due to true differences and $SS_{(within)}$ the sum of squares due to error. An estimate of the reliability coefficient could, therefore, be obtained from :

$$\frac{SS_{(between)}}{SS_{(total)}} = \eta^2 = r_{xx}$$

However, more sophisticated methods which estimate the value of true and error variance more accurately are preferred. One of these is the intra-class correlation coefficient. This is derived from the assumptions of the model involved in the analysis of variance. By the assumptions of the model (random effects or Model 2):

(1) $$SS_{(within)} / n - 1 = MS_{(between)} = \sigma_e^2 + m\sigma_B^2$$ (12: 18)

and

(2) $$SS_{(within)} / nm - n = MS_{(within)} = \sigma_e^2$$ (12:19)

where:

n = number of groups, i.e. in this case, number of individuals tested

m = number within groups, i.e. number of times each individual is tested

MS = mean square or population variance estimate

σ_e^2 = error variance

σ_B^2 = variance due to non-chance differences between groups,
i.e. in this case, true score variance

Reliability is defined as the ratio of true variance to total variance so, as total variance will equal true variance plus error variance, it is necessary to form the ratio :

$$\frac{\sigma_B^2}{\sigma_B^2 + \sigma_e^2}$$ (12:20)

This can be done by using the following formula:

$$\frac{MS_{(between)} - MS_{(within)}}{MS_{(between)} + (m-1)MS_{(within)}} = r_{xx}$$ (12:21)

Proof

(1) Using (12:18) and (12:19), (12:21) can be written as:

$$\frac{m\sigma_B^2 + \sigma_e^2 - \sigma_e^2}{m\sigma_B^2 + \sigma_e^2 - (m-1)\sigma_e^2}$$

(2) This simplifies to:

$$\frac{m\sigma_B^2}{m\sigma_B^2 + m\sigma_e^2}$$

(3) Dividing by *m* gives:

$$\frac{\sigma_B^2}{\sigma_B^2 + \sigma_e^2} = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_e^2} = r_{xx}$$

Although this discussion has been in terms of true scores and error scores it is interesting to note that (12:21) approximates to the mean intercorrelation between tests. The proof of this previous statement would take us too far afield, but it will be recalled that the mean intercorrelation between items or tests is related to reliability in the domain sampling model.

For further details of the assessment of reliability by analysis of variance techniques see McNemar (1969) and Winer (1962).

The assessment of individual results

1. *Introduction*

This chapter will be concerned with a number of topics more or less directly related to what has gone before. It should be emphasised that a large number of statistical methods can be applied to the study of individual cases and references to these will be found at the end of the book. We will be concerned only with the reliability of differences in score and changes in score, and the assessment of the rarity or abnormality of differences or changes in score. Specifically we will deal with the following questions:

(1) Is a difference in score between two individuals on the same test likely to be due to chance?

(2) Is a difference in score for the same individual on two occasions likely to be due to chance?

(3) Is a difference between an individual's scores on two tests likely to be due to chance?

(4) How rare is a given difference between an individual's scores on two tests? Is it abnormally large?

(5) How rare is a given change in an individual's score in the same test? Is it large enough to be considered abnormal?

2. *Differences between Two Individuals on the Same Test*

The problem here is to decide how likely it is that two different obtained scores represent two different true scores. If we had the standard deviation of the distribution of differences between obtained scores when the true scores are the same, we could work out a *Z* score for the difference we obtain and look this up in tables for the normal curve. We could then see what proportion of differences, when true scores do not differ, would be larger than the one we have obtained. This proportion would give us the probability that the two obtained scores in fact represent two identical true scores. It is not too difficult to work out what the distribution should be. We will use deviation scores to make the derivation simpler. σ^2_{diff} is the variance of the difference between scores :

(1)
$$\sigma^2_{diff} = \frac{\sum\left[(x_1 - x_2) - (\bar{x}_1 - \bar{x}_2)\right]^2}{N}$$

(2) but $x_1 = (t_1 + e_1);$ and $x_2 = (t_2 = e_2)$ and $\bar{x}_1$ and $\bar{x}_2$ are mean deviation scores and thus equal to zero. So

$$\sigma^2_{diff} = \frac{\sum\left[(t_1 + e_1) - (t_2 + e_2)\right]^2}{N}$$

(3) But we are concerned with the variance of differences between scores when *ti* = *t*$_2$. So:

$$\sigma^2_{diff} = \frac{\sum\left[(t + e_1) - (t + e_2)\right]^2}{N}$$

$$= \frac{\sum(e_1 - e_2)^2}{N}$$

(4) This equals

$$\frac{\sum\left(e_1^2 + e_2^2 - 2e_1e_2\right)}{N}$$

(5) $$= \frac{\sum e_1^2}{N} + \frac{\sum e_2^2}{N} - 2\frac{\sum e_1e_2}{N}$$

(6) But the last term is the covariance of error scores and equals zero and the first terms are error variances so

$$\sigma_{diff}^2 = \sigma_{e_1}^2 = \sigma_{e_2}^2$$

(7) But the error variances will also be equal and will equal σ_{meas} squared, so:

$$\sigma_{diff}^2 = 2\sigma^2\left(1 - r_{xx}\right)$$

(8) The square root of this will be the standard deviation of the distribution of differences between obtained scores when the true scores are identical.

$$\sigma_{diff} = \sqrt{2\sigma^2\left(1 - r_{xx}\right)} = \sigma_{meas}\sqrt{2} \qquad (13:1)$$

Thus to see whether two obtained scores differ, their differences should be divided by $\sigma_{meas}\sqrt{2}$. This will yield a Z score for the distribution of differences obtained on a chance basis, and by reference to tables for the normal curve the significance of this difference can be assessed.

Problem

A obtains a score of 90 on a test with a mean of 100 and a standard deviation of 10. On the same test *B* obtains a score of 104. If the reliability coefficient of the test is .755, with what degree of certainty can we conclude that *B*'s score is really higher than *A*'s?

Answer

The $\sigma_{diff} = \sqrt{2\sigma^2(1 - r_{xx})} = \sqrt{2 \times 100 \times .245} = 7.$ The difference between $A's$ score and *B*'s score is 14 points. Dividing this by 7 gives an answer of 2. A difference of 14 points is, therefore, 2 standard deviations away from the mean of differences obtained on a chance basis. Less than 5 per cent of chance differences will be as large as this.

3. *Differences between Scores on the Same Test for the Same Individual on Two Occasions*

The problem here is to find whether a change in scores obtained on two separate occasions is likely to be due to chance. Once more what is needed is the distribution of differences between two obtained scores when the true scores are in fact the same. This is precisely the same situation as the one above, and the appropriate *Z* value can be found by:

$$\frac{\text{Difference between scores}}{\sigma_{meas}\sqrt{2}}$$

One slight problem here is that if the individual takes the same test twice there are almost certain to be practice effects. If these are known the formula is modified to take account of them thus:

$$\frac{\text{Difference between scores - practice effect}}{\sigma_{meas}\sqrt{2}}$$

4. *Differences between Scores on Two Different Tests/or One Individual*

In this case the distribution of interest is the distribution of differences between obtained scores on two different tests when the scores on each test are in fact the same. For this to be a sensible procedure the scores on each test should be in comparable units, e.g. *T.* scores, I.Q.s with same mean and *σ*, or *Z* scores, because we are not interested in differences in the scores as such, but in differences in the individuals' relative standing on the two tests. The derivation of the formula follows the same steps as those above, except that we have *x* and *y* as our deviation scores.

(1) $$\sigma^2_{diff} = \frac{\sum\left(x - y^2\right)}{N}$$

(2) $$= \frac{\sum\left[\left(t_x + e_x\right) - \left(t_y + e_y\right)\right]^2}{N}$$

(3) Because we are interested in the situation where $t_x = t_y$ this becomes:

$$\frac{\sum\left(e_x^2 + e_y^2 - 2e_x e_y\right)}{N}$$

(4) This equals $\sigma^2_{e_x} + \sigma^2_{e_y}$

(5) So $$\sigma^2_{diff} = \sigma^2_{e_x} + \sigma^2_{e_y} = \sigma^2_{meas.x} + \sigma^2_{meas.y}.$$

If this is computed in *Z* scores the standard deviation will be 1, so:

$$\sigma_{meas.x} + \sigma_{meas.y} = \sigma_x\sqrt{1-r_{xx}} + \sigma_y\sqrt{1-r_{yy}}$$

$$= \text{ (in Z score terms)}\sqrt{2-\left(r_{xx}+r_{yy}\right)}$$

So to assess the probability that a difference between scores on two tests is due to chance we have the formula:

$$\frac{Z_x - Z_y}{\sqrt{2-\left(r_{xx}+r_{yy}\right)}} \qquad (13:2)$$

Problem

On a test of intelligence involving visual material an individual scores at the 75th percentile, while on a test involving verbal material the score is at the 50th percentile. Is there any reason for supposing that the difference between the test materials is affecting the individual's performance, if the reliabilities of the tests are .80 and .84 respectively?

Answer

The problem here is whether this difference is attributable to chance or not. The formula is:

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

Filling in the appropriate values we obtain:

$$\frac{.67 - .00}{\sqrt{2 - 1.64}} = \frac{.67}{.60} = 1.12$$

The difference is, therefore, not very large, and by chance approximately 26 per cent of differences would be larger than this.

5. *Abnormality of a Difference between Scores on Different Tests*

So far we have been concerned with the problem of whether differences observed are due to errors of measurement. In this and the next section our concern will be with whether a difference is abnormally large in the sense that it is so large as to happen rarely. For this purpose we need the standard deviation of the distribution of differences between scores.
The derivation is as follows:

(1) $$\sigma^2_{diff} = \frac{\sum\left[(X-Y)-(\overline{X}-\overline{Y})\right]^2}{N}$$

(2) $$= \sum\left(X^2 + Y^2 + \overline{X}^2 + \overline{Y}^2 - 2XY - 2X\overline{X} + 2X\overline{Y} + \right)$$

(3) $$= \frac{\sum X^2}{N} + \frac{\sum Y^2}{N} + \frac{N\overline{X}^2}{N} + \frac{N\overline{Y}^2}{N} - \frac{2\sum XY}{N} - \frac{2\sum X}{N}\overline{X}$$

$$+2\frac{\sum X}{N}Y + 2\frac{\sum Y}{N}\overline{X} - 2\frac{\sum Y}{N}\overline{Y} - 2\frac{N\overline{X}\overline{Y}}{N}$$

(4) But, $2\frac{\sum X}{N}\overline{Y} = 2\overline{X}\overline{Y}, 2\frac{\sum X}{N}\overline{X} = 2\overline{X}^2$ and so on, so we obtain:

$$\frac{\sum X^2}{N} - \overline{X}^2 + \frac{\sum Y^2}{N} - \overline{Y}^2 - 2\left(\frac{\sum XY}{N} - \overline{X}\overline{Y}\right)$$

(5) This equals:

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

(the last term is a covariance term)

(6) The square root of this will be the standard deviation of differences.

$$\sigma_{diff} = \sqrt{\sigma_x^2 + \sigma_y^2 - 2r_{xy}\sigma_x\sigma_y}$$

In terms of Z score the formula becomes:

$$\frac{Z_x - Z_y}{\sqrt{2 - 2r_{xy}}} \qquad (13{:}3)$$

Once more a raw score version should only be used when scores on both tests have the same mean and standard deviation.

Problem

On a test with a mean of 100 and standard deviation of 15 an individual obtains a score of 130. On a second test with a mean of 100 and a standard deviation of 10 he scores 90. If the correlation between these tests is .50, how abnormal is this discrepancy?

Answer

Formula (13:3) is appropriate here:

$$\frac{Z_x - Z_y}{\sqrt{2 - 2r_{xy}}} = \frac{2.00 - (-1.0)}{\sqrt{2 - (2 \times .50)}} = 3.0$$

The difference is therefore 3 standard deviations away from the mean of differences expected on the basis of the correlation between the tests. Only about 3 cases in 1,000 would show a difference as large as this.

6. *The Abnormality of a Change in Score*

A method for assessing changes in score for an individual, when expected change is due only to the unreliability of the test or measure has been described in Section 3 of this chapter. Sometimes, however, expected change is due to factors other than poor reliability. Over a period of time real change might have occurred. Degree of depression, amount of anxiety, a level of skilled performance are examples of variables which might show real change with passing time. In this kind of situation two distinct questions can be asked about the observed change:

(1) is it so great that it is unlikely to be due to errors of measurement—the problem dealt with in Section 3; and

(2) is it so great that it is unlikely to be due to errors of measurement and normal real changes?

The method for dealing with the second problem involves:

(1) using a regression equation to predict the second score from the first;

(2) finding the difference between the obtained and predicted scores;

(3) assessing the significance of this difference in terms of the standard error of estimate.

This can be done using the formulae described in Chapters 5 and 6. The score on the first occasion will be called X and the score on the second occasion will be called Y, and the formula will be presented in Z score terms, r is the correlation between test and retest over the period of interest.

(1) $\hat{Z}_y = r_{xy}Z_x$ = predicted score;

(2) $\hat{Z}_y - Z_y$ = difference between obtained and predicted score;

(3) the standard error of estimate is (in Z score terms):

$$\sqrt{1 - r_{xy}^2}$$

So the *Z* score for the distribution of obtained scores around predicted scores will be:

$$\frac{Z_Y - \hat{Z}_Y}{\sqrt{1 - r_{xy}^2}} \qquad (13{:}4)$$

Once more the situation can be complicated by practice effects. To allow for these $\hat{Z}_Y$ can be modified to equal:

$$r_{xy}Z_X + \frac{\text{practice effects}}{\sigma_{\mathrm{y}}}$$

If the *Z* obtained by use of (13:4) is converted to a percentile it will give the percentage of people, starting with the same score as the individual of interest, who would be expected to obtain a lower (or higher) score than him on the second occasion.

Problems

A. A child is tested at the age of 3 years on a test of intelligence and obtains a score of 160, and retested at age 7 obtaining a score of 155. If the test-retest correlation over this period of time is .50, and the mean and standard deviation are 100 and 15 respectively, what proportion of children would show a drop in score as large as this. (Assume that practice effects are negligible over this period of time.)

B. After relaxation training of three months' duration a patient's anxiety score drops from the 84th to the 50th percentile. The therapist concludes that a change of this magnitude must indicate that relaxation has been successful in treating the patient's anxiety. If for untreated cases the test-retest correlation of the anxiety measure (over 3 months) is .60, is the therapist justified?

Answers

A. Inserting values in (13:4) gives:

$$\frac{3.67-(4.00\times.50)}{\sqrt{1-.50^2}}=\frac{1.67}{.87}=1.92$$

The difference of minus 5 points is thus 1.92 standard deviations *above* the mean of expected differences. Approximately 97 per cent of children would show a greater drop. Less than 5 per cent of children who started with a score of 160 at age 3 would be expected to obtain a score as high as 155 at age 7. It might, therefore, be worth considering factors which might have caused a *rise* in intelligence in this child.

B. The change is not so large as to be very rare. Using formula (13:4) again gives:

$$\frac{0-(.60\times1.0)}{\sqrt{1-.60^2}}=\frac{-.60}{.80}=-.75$$

The second anxiety score is three quarters of a standard deviation below what would be expected, but about 23 per cent of untreated patients with the patient's initial score would be expected to show a greater drop in anxiety in the normal course of events.

Chapter Fourteen

Classification

1. *Introduction*

This chapter is concerned with a number of problems which arise in the use of tests for classification purposes. These are:

(1) The selection of cut-off points.
(2) The effects of base rates on the usefulness of tests.
(3) The effects of selection ratios on the effectiveness of selection procedures.

If we are concerned with the prediction of continuous variables, and are interested only in the value of scores on them, none of the above problems arise, but most practical uses of tests are in fact concerned with allocating subjects to groups. The problems are therefore real ones to the test user and an introductory account is thus desirable.

2. *Selection of Cut-off Points*

A problem which frequently arises in the use of tests is the problem of locating a cut-off score for use in differentiating two groups. It would be nice if distributions of scores of different groups were separate from one another but in real life they inevitably overlap. The scores of schizophrenics on tests of thought disorder overlap with those of neurotics, the scores of successful salesmen on intelligence and interest tests overlap with those of unsuccessful salesmen, and so on. Such a situation is shown in Figure 14.1 for two groups *A* and *B*.

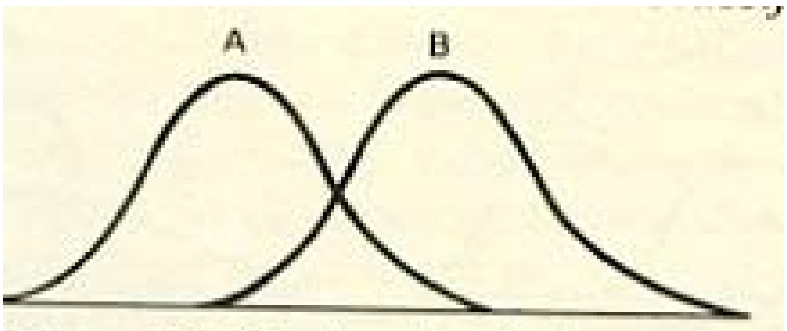


Figure 14.1 Overlapping distributions

The problem is to find that point which classifies *A* and *B* as accurately as possible, if we make the rule that subjects scoring on one side of the point will be called A and those on the other called B. In fact the solution is quite simple. If we take as our cut-off point the point where the distributions overlap, we will have achieved the best discrimination possible. This is shown in Figure 14.2.

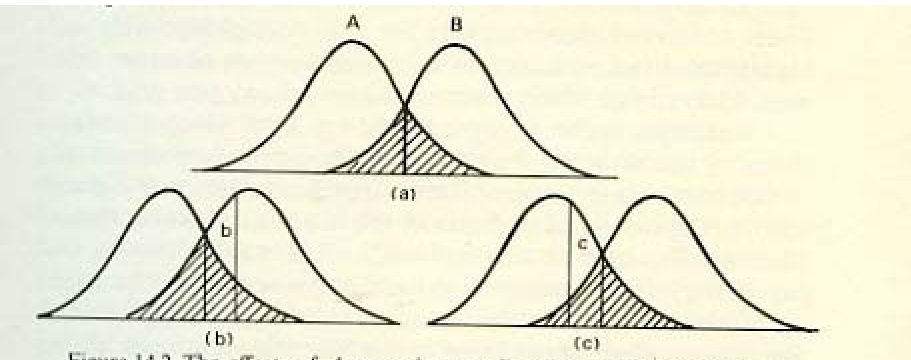


Figure 14.2 The effects of changes in cut-off point on misclassification

In (a) the cut-off point is at the point of intersection of the two curves and the cases misclassified will fall in the shaded area. In (b) the cut-off point is moved to a higher value. The original area of misclassification is still there, shaded, but it has now had added to it, the cases in area *b*. In (c), the cut-off point has been set at a lower value. Again the original amount of misclassification is obtained plus the cases in area *c*. So it can be seen that putting the cut-off point at any point other than the intersection of the curves, will lead to an increase in errors of classification. Note that the point where the curves intersect will be the point where the number of cases in *A* equals the number of cases in *B*, and this fact can be used to find the best cut-off point.

Problem
Where would you place cut-off points in Figure 14.3 *a* and *b*?

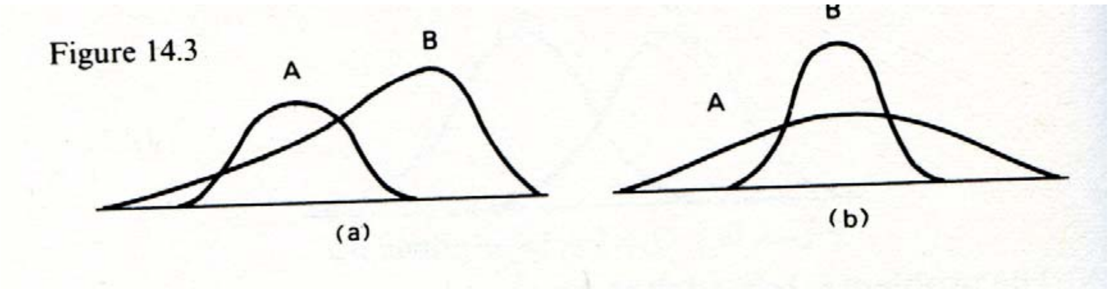


Figure 14.3

Answer

> The solution involves the placing of two cut-offs. In each case a cut-off is placed at the points where the curves intersect; Placing at any other point will lead to a greater number of errors.

3. *The Effects of Base Rates on the Usefulness of Tests*

There are circumstances where the use of a highly reliable and highly valid test will lead to a greater number of errors than would have been made without the use of any test at all.

A base rate is the relative frequency with which a certain category or characteristic appears in the population of interest. The base rate for schizophrenia in a mental hospital is given by the proportion of patients in the hospital who are schizophrenic. The base rate for schizophrenia in cases seen by the psychology department will in most cases be different from the hospital base rate. Only if the cases referred are a representative sample of all of the cases in the hospital will the two base rates be the same. Similarly the base rate for successful executives in a given industry may well be different from the base rate for successful executives amongst applicants to a management selection firm selecting for that industry.

Given the base rates for various categories in the population of interest, what is the best classification rule to follow in the absence of any other information? To make this concrete let us suppose that amongst cases referred to a psychology department in a hospital 20 per cent are brain damaged, 35 per cent schizophrenic, 25 per cent affective psychosis, and 20 per cent neurotic. What rule can be applied in this situation to make the smallest number of errors?

If we call every patient brain damaged we will be wrong in 80 per cent of cases, if we call every patient schizophrenic we will be wrong in 65 per cent of cases, if we call everyone an affective psychotic, we will be wrong in 75 per cent and if everyone is called neurotic 80 per cent will be misclassified. Clearly the smallest number of errors occurs when we call everyone schizophrenic. This leads to only 65 per cent error. Any other decision will increase errors over this figure. So by choosing the commonest diagnosis – schizophrenia - we make fewer mistakes than by choosing any other diagnosis.

The general principle is that, in the absence of other information, fewest errors will be made by guessing that everyone is in the category with the highest base
rate.

Sometimes accuracy using base rates alone can be very high. Suppose that the problem in the example above had been to classify patients as brain damaged or not brain damaged. If we called every patient brain damaged we would be wrong in 80 per cent of cases. If we called everyone not brain damaged, then we would be wrong in only 20 per cent of cases.

We now need to introduce the concept of conditional probability. Continuing with our example let us suppose that in the hospital from which the psychology department's population is drawn the base rate for brain damage is 5 per cent. Thus the probability that a patient drawn at random from this population will be brain damaged is $pBD = .05$. The probability that the patient is brain damaged given that he has been referred to the psychology department is $pBD/RTPD$—probability of brain damaged given that (/) the patient has been referred to the psychology department is .20. This is known as a conditional probability. Generally pA/B can be read as the probability of A given that B occurs, or the probability of A conditional on B. The standardisation data of diagnostic tests and classification instruments are often given in terms of the proportions of a group obtaining a score above and below a criterion. These data can be worked on as conditional probabilities. Consider the following hypothetical example. A test of brain damage correctly classifies 60 per cent of brain-damaged patients, and 90 per cent of not-brain-damaged patients. The data are shown below:

		Test Diagnosis	
		Brain Damage	*No Brain Damage*
True Diagnosis	*Brain Damage*	60 per cent	40 per cent
	No Brain Damage	10 per cent	90 per cent

The rows of the table give the probability of a brain damage score conditional on a given true diagnosis.

$$pTBD/BD = .60;\ pTNBD/BD = .40 \text{ and}$$

$$pTBD/NBD = .10;\ pTNBD/NBD = .90$$

where

TBD = test diagnosis of brain damage
BD = true diagnosis of brain damage
NBD = true diagnosis of no brain damage
$TNBD$ = test diagnosis of no brain damage

However, when we use the test we are interested in probabilities which are conditional in the other direction. We do not want to know *pTBD*/*BD* or *pTBD*/*NBD*, we want to know p*BD*/*TBD* and *pNBD*/*TBD* and so on. Given a test score we want to know the probability of the diagnosis. On reflection we can see that the proportion of the population who will be both (a) brain damaged and (b) have a test score indicating brain damage will be:

Proportion of brain damaged in population

Multiplied by

Proportion of brain damaged
who obtain brain damaged score on test

The proportion of brain damaged patients is of course the base rate for brain damage. So *p*(*BD* and *TBD*) = *pBD* x *pTBD*/*BD*. However the proportion of patients who are both (a) brain damaged and (b) obtain a brain damaged score is only half the story. Other people will also get brain damaged scores. If we want to know the proportion of people obtaining brain damaged scores who are in fact brain damaged we need to work out the ratio.

Proportion who are:
(a) brain damaged, and
(b) get brain damaged scores

Divided by

Total proportion of patients
who get brain damaged scores

Fortunately it is easy to work out the proportion of patients who are both (a) not-brain-damaged and (b) who get brain damaged scores. It will be

Proportion of not-brain-damaged patients

Multiplied by

Proportion of not-brain-damaged patients
who obtain brain damaged scores.

This equals:

$$pPNBD \times pTBD/NBD$$

where $pNBD$ is the base rate for no brain damage.

Putting this together we obtain:

$$pBD/TBD = \frac{pBD \times pTBD/BD}{(pBD \times pTBD/BD) + (pNBD \times pTBD/NBD)} \qquad (14:1)$$

Similarly it can be shown that:

$$pNBD/TNBD = \frac{pNBD \times pTNBD/NBD}{pNBD \times pTNBD/NBD) + (pBD \times pTNBD/BD)} \qquad (14:2)$$

The proportion wrongly classified will consist of:

(1) Patients who are
(a) brain damaged, and
(b) obtain a normal score
plus

(2) Patients who are
(a) not brain damaged, and
(b) obtain a brain damaged score

The first value will be given by:

proportion of brain damaged patients x proportion of brain damaged who get normal score = $pBD \times pTNBD/BD$.

and the second by

> proportion of non brain damaged patients x proportion of non-brain-damaged who obtain brain damaged scores

$$= \quad pNBD \text{ x } \mathrm{p}TBD/NBD.$$

As an example of the use of these formulae let us suppose that the test, whose standardization data were given above (and are repeated here) is used on a population where the base rate for brain damage is 20 per cent. The base rate for non-brain damge will thus be .80.

		Test Diagnosis	
		Brain Damage	*No Brain Damage*
True Diagnosis	*Brain Damage*	60 per cent	40 per cent
	No Brain Damage	10 per cent	90 per cent

Using the formulae we can see:

(1) that the probability of the patient being brain damaged if he gets a brain damaged score = pBD/TBD =

$$\frac{.20 \times .60}{(.20 \times .60) + (.80 \times .10)} = \frac{.12}{.20} = .60$$

(2) that the probability of the patient being not brain damaged if he gets a not brain damaged score = $pNBD/TNBD$ =

$$\frac{.80 \times .90}{(.80 \times .90) + (.20 \times .40)} = \frac{.72}{.80} = .90$$

(3) that the total proportion misclassified will be

$$pBD/TNBD \quad + pNBD/TBD$$

$$= (.20 \text{ x } .40) \;+ (.80 \text{ x } .10) = .16$$

(4) the proportion correctly classified will be

1 — proportion misclassified = 1 — .16 = .84

Just using the base rate we would have guessed that any patient from this population was not brain damaged and we would have been right 80 per cent of the time. The use of the test has therefore increased our proportion of successes from .80 to .84.

If the base rate for brain damage had been .l0 instead of .20, the picture would have been as follows:

(1) $pBD/TBD = \dfrac{.10 \times .60}{(.10 \times .60) + (.90 \times .10)} = \dfrac{.06}{.15} = .40$

So the majority of patients with a brain damage score would have been not brain damaged.

(2) $pNBD/TNBD = \dfrac{.90 \times .90}{(.90 \times .90) + (.10 \times .40)} = \dfrac{.81}{.85} = .95$

(3) $pBD/TNBD \quad + \quad pNBD/TBD =$

$$(.10 \text{ x } .40) + (.90 \text{ x } .10) = .13$$

In this case use of the base rate alone would have lead to only 10 per cent errors, while using the test has increased this to 13 per cent. So we have made more mistakes with the test than without it.

Problems

Given the following data:

		Test Diagnosis	
		Brain Damage	*Normal*
True Diagnosis	*Brain Damage*	.50	.50
	Normal	.20	.80

A. What is the probability that a patient who obtains a brain damaged score is brain damaged in a population where the base rate for brain damage is .20, and for normalcy .80?

B. What is the probability that someone obtaining a normal score is brain damaged?

Answers

A. pBD/TBD

$$= \frac{pBD \times pTBD/BD}{(pBD \times pTBD/BD) + (pNBD \times pTBD/NBD)}$$

$$= \frac{.20 \times .50}{(.20 \times .50) + (.80 \times .20)} = .38$$

B. $pBD/TNBD$

$$= \frac{pBD \times pTNBD/BD}{(pBD \times pTNBD/BD) + (pNBD \times pTNBD/NBD)}$$

$$= \frac{.20 \times .50}{(.20 \times .50) + (.80 \times .80)} = .135$$

The methods can be easily extended to more than two groups. Suppose that the following data are available.

		Test Diagnosis	
		Brain Damage	*Not Brain Damaged*
True	*Brain Damage*	60	40
Diagnosis	*Psychosis*	30	70
	Neurosis	20	80
	Normal	10	90

Given base rates of:

Brain damage	.10
Psychosis	.40
Neurosis	.40
Normalcy	.10

The probability of a diagnosis of brain damage given a test score indicating brain damage will be pBD/TBD

$$= \frac{pBD \text{ x } \mathrm{p}TBD/BD}{(pBD \text{ x } pTBD/BD) + (p\mathrm{Psychosis} \text{ x } pTBD/\mathrm{Psychosis}) + (p\mathrm{Neurosis} \text{ x } pTBD/\mathrm{Neurosis}) + (p\mathrm{Normalcy} \text{ x } pTBD/\mathrm{Normalcy})}$$

$$= \frac{.10 \times .60}{(.10 \times .60) + (.40 \times .30) + (.40 \times .20) + (.10 \times .10)}$$

$$= \frac{.06}{.27} = .22$$

If we repeat this calculation for all groups and work out pPsychosis/*TBD*, pNeurosis/*TBD*, and pNormal/*TBD* we will obtain the following values:

*p*Psychosis/*TBD* = .44
*p*Neurosis/*TBD* = .29
*p*Normalcy/*TBD* = .04

Thus the most frequent true diagnosis amongst those diagnosed by the test as brain damaged will be psychosis, If we bet that everyone with a brain damage score is psychotic we will make fewer mistakes than if we use other diagnosis. All of this may lead to gloomy thoughts about the value of tests but it is worth noting that if in these examples we had guessed that everyone was brain damaged, only 10 percent would have been, whereas if we call all of those getting a test score indicating brain damage brain damaged, 22 per cent will in fact be brain damaged. In this sense the test has been useful. Further a test diagnosis of not brain damaged is pretty useful. The proportion not brain damaged when the test says no brain damage will be

((*p*Psychosis x *pTNBD*/Psychosis)
+ (*p*Neurosis x *pTNBD*/Neurosis)
+ (*p*Normalcy x *p*TN *BD*/Normalcy))

divided by

((*p*Psychosis x *pTNBD*/Psychosis

+ (*p*Neurosis x *p*TNB/Neurosis)

+ (*p*Normalcy x *pTNBD*/Normalcy)
+ (*pBD* x *pTNBD*/*BD*))

$$= \frac{(.40 \times .70) + (.40 \times .80) + (.10 \times .90)}{(.40 \times .70) + (.40 \times .80) + (.10 \times .90) + (.10 \times .40)}$$

$$\frac{.69}{.73} = .95$$

Problems

(1) If in the above example the base rates had been:

(a) Brain damage	.40
(b) Psychosis	.20
(c) Neurosis	.20
(d) Normalcy	.20

work out:

(i) the value of pBD/TBD
(ii) the value of pNormal/TBD
(iii) the value of $pNBD/TNBD$
(iv) the value of pPsychosis/$TNBD$

(2) It is probably important to detect as many brain damaged patients as possible. Suppose the cut-off were moved so as to identify a greater proportion of the brain damaged, what would happen to the value of $PNBD/TBD$?

Answers

(1) (i)
$$\frac{(.40\times.60)}{(.40\times.60)+(.20\times.30)+(.20\times.20)+(.20\times.10)}$$
$$=\frac{.24}{.36}=.67$$

(ii)
$$\frac{(.20\times.10)}{(.40\times.60)+(.20\times.30)+(.20\times.20)+(.20\times.10)}$$

$$=\frac{.02}{.36}=.06$$

(iii) $$\frac{(.20\times.70)+(.20\times.80)+(.20\times.90)}{(.20\times.70)+(.20\times.80)+(.20\times.90)+(.40\times.40)}$$

$$=\frac{.48}{.64}=.75$$

(iv) $$\frac{(.20\times.70)}{(.20\times.70)+(.20\times.80)+(.20\times.90)+(.40\times.40)}$$

$$=\frac{.14}{.64}=.22$$

3. *The Effects of Selection Ratios on the Effectiveness of Selection Procedures*

A selection ratio is usually defined as the number of jobs available divided by the number of applicants for those jobs. If there are five applicants available for every job the selection ratio is .20; if there are twenty applicants for every job it is .05 and so on. Given a valid predictor of job success and a selection ratio of .20, one would give the jobs to the 20 per cent with the highest scores on the predictor, or more generally give the jobs to the lop *N* per cent of applicants (where *N* = selection ratio x 100). The effectiveness of this in terms of the proportion selected who are satisfactory varies with:

(1) the size of the correlation between predictor and criterion
(2) the selection ratio
(3) the criterion cut-off point, from which can be deduced the proportion who are potentially successful, i.e. the base rate for success amongst applicants

The reader will not be surprised to hear that the proportion of those selected who are judged successful depends on the size of the correlation coefficient, because all we are saying is that prediction gets better as r_{xy} gets higher. The other relationships may not be quite so obvious. Figure 14.4 shows three scatter diagrams, a, b, and c.

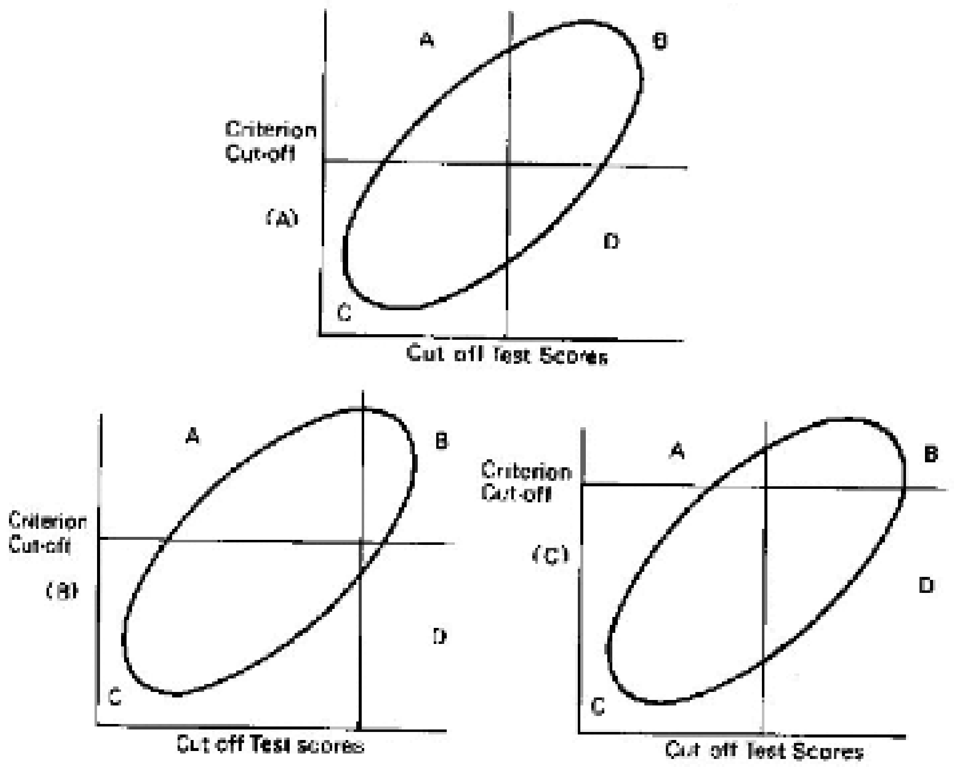


Figure 14.4 Successes (*B* and *C*) and errors. (*A* and *D*) in relation to criterion and predictor (test scores) cut-off points

In (a) the selection ratio is approximately .50 and the proportion above the criterion cut-off is also about .50. If we use our test for selection we will select applicants filling in areas *B* and *D*, and of these a proportion $\frac{B}{B+D}$ will be successful, and the rest not successful.

In scatter diagram (b) the criterion cut-off remains the same but the selection ratio has decreased. Fewer applicants are now selected. Once more the applicants selected will be those in areas *B* and *D* but it can be seen that the proportion of successes $\frac{B}{B+D}$ is now much larger. So as the selection ratio *decreases* so the proportion of those selected who are successes *increases.* Now look at the proportion of cases in area *A*. These are those who would have been successful who have been rejected by the test. As the selection ratio decreases the ratio of this group to those selected goes up. So a decrease in the selection ratio leads to:

(1) a higher proportion of satisfactory employees amongst those selected
(2) a smaller proportion of all potentially satisfactory employees being selected

In the industrial situation one is more concerned with the first of these, but for educational purposes, or for purposes of selecting patients for treatment the second is also important.

In diagram (c) the selection ratio is once more approximately .50, but this time the criterion cut-off has been raised. Now a smaller proportion of employees is considered satisfactory. The effect of this is to decrease the ratio *B/(B + D),* and also to decrease the ratio *A/(A + B)* so raising the criterion cut-off leads to:

(1) a decrease in the proportion of those selected who are considered satisfactory
(2) an increase in the proportion of those potentially satisfactory who are selected

The three variables, selection ratio, proportion considered satisfactory, and the size of the correlation coefficient all influence the size of the proportion of those selected who are considered satisfactory. Taylor and Russell (1939) have prepared tables which enable one to obtain the proportion correctly selected from knowledge of the three variables. Parts of their tables are reproduced below in Table 14:1.

TABLE 14:1 SELECTED VALUES FROM THE TAYL OR-RUSSELL TABLES

(a) Proportion considered satisfactory = .20

		Selection Ratio				
		.10	***.30***	***.50***	***.80***	***.90***
	.10	.25	.23	.22	.21	.21
r	***.30***	.37	.30	.27	.23	21
	.50	.52	.38	.31	.24	.22
	.80	.79	.53	.38	.25	.22
	.90	.91	.60	.40	.25	.22

(b) Proportion considered satisfactory = .80

		Selection Ratio				
		.10	***.30***	***.50***	***.80***	***.90***
	.10	.85	.83	.82	.81	.81
r	***.30***	.92	.89	.87	.84	.82
	.50	.97	.94	.91	.86	.84
	.80	1.0	1.0	.98	.91	.87
	.90	1.0	1.0	1.0	.94	.88

Each table shows the effects of increasing r_{xy} and increasing the selection ratio. The effects of changing the criterion cut-off can be assessed by comparing the tables. It is worth noting that the original tables give much more detail than is given here. All values of r_{xy} from 0 to 1.0 in steps of .05 r_{xy} are given, as are the values for selection ratios of .05 to .95, and values of proportion of employees considered satisfactory from .05 to .90

.

Let us now see how we could use these tables in a clinical situation. Suppose that a researcher has found a significant difference in Neuroticism scores between phobics cured or much improved by systematic desensitization, and those not so helped. Let us further suppose that in this investigation the ratio for the difference between means was 3.0 and that the number of patients involved was 93. Using formula (12:7) we can convert this value to a measure of association thus:

$$\text{eta} = \sqrt{\frac{t^2}{t^2 + df}} = \sqrt{\frac{9}{9+91}} = \sqrt{.09} = .30$$

If we can assume a linear relationship between *N* scores and degree of improvement we use the information in the Taylor-Russell Tables to find out what proportion of those selected for treatment will in fact benefit. Suppose that we use *N* scores to select, and that we accept 50 per cent of those referred, and that the success rate for systematic desensitization in phobics is 80 per cent. (This last figure is analogous to the proportion considered satisfactory.)

We can now use Table 14: 1(b) to see the proportion of those selected who will be cured or much improved by treatment. This value turns out to be .87. Using no selection at all we would have cured 80 per cent of those treated, using *N* scores to select we cure 87 per cent of those treated. But as this is a clinical situation other questions must be asked.

(1) What proportion of those rejected could have been cured by treatment?

(2) What proportion of those who could have been cured have been cured?

Question (1) can be answered as follows. We know that 50 percent of the patients were selected for treatment, and that of this 50 per cent 87 per cent were cured. So the proportion of the total population who were (a) selected and (b) cured will be .50 x .87 = .435. Further we know that .80 of the total population could have been cured, so the proportion of those rejected who could have been cured will be given by

$$\frac{.80-.435}{.50}=.73$$

So the use of our selection procedure has lead to the rejection of a group of patients 73 per cent of whom could have been cured by treatment.

The second question was concerned with the proportion of those who could have been cured who have been cured. This will be the proportion cured, over the proportion who could have been, or:

$$\frac{.435}{.80}=.545$$

So the use of the selection procedure leads to the cure of 54.5 per cent of those who might have been cured, while the use of no selection procedure would have lead to the cure of 100 per cent of curable patients.

It will be objected that no-one ever chooses patients for treatment on the basis of one characteristic. This may be true, but the only difference it makes is that we need the validity coefficient for the selection procedure as a whole instead of just for one measure. The moral of this example is clear. Selection for treatment in clinical situations must take into account the factors outlined above. Otherwise clinicians may be doing their patients a grave injustice.

A. If in the above example the success rate for treatment had been 20 per cent:

(a) what proportion of those selected would have been cured?

(b) what proportion of all who could have been cured would have been cured?

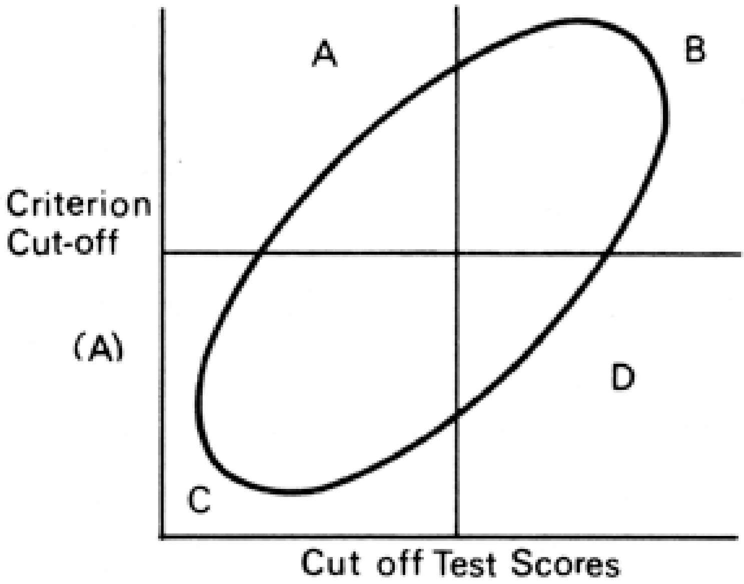
B. Referring to the diagram below and using the areas *A, B, C,* and *D,* indicate the areas included by:

(a) total considered satisfactory

(b) proportion of successes amongst those selected

(c) proportion of potential successes selected

(d) proportion of failures correctly classified



Answers

A. (a) Referring to Table 14: l(a), with a success rate of .20, a selection ratio of .50 and a correlation between predictor and criterion of .30, the percentage of those selected who will be cured is 27 per cent.

(b) 13.5 per cent of the sample will be (1) selected and (2) cured. Altogether 20 per cent are curable, therefore $\frac{13.5}{20.0}$ or 67.5 per cent of those curable will have been cured.

B. (a) $A + B$

(b) $\frac{B}{B+D}$

(c) $\frac{B}{A+B}$

(d) $\frac{C}{C+D}$