

RELIABILITY

OBJECTIVE:

To increase understanding of the concept of reliability and to provide first hand experience with the calculation of interscorer and test-retest reliabilities.

GENERAL INFORMATION - RELIABILITY

The term reliability as used in the field of psychological testing refers to **consistency** in measurement. Classical Test theory or "true score" theory proposes that a score on a test consists of two parts. The first part is the "true" amount of whatever is being measured (e.g., intellectual ability on an intelligence test, or depression on a depression scale). The second part consists of unsystematic error of measurement (e.g. time of day the test was given, the particular sample of questions chosen for the test, the lighting in the room, a faulty pen, etc.). All of these extraneous variables may affect test takers differentially. This part of the score is referred to as the error portion of the score.

A reliability coefficient is a numerical index of reliability that expresses the ratio between the "true" score on a test and the total variance and basically gives an indication of how free the test is from errors in measurement. In test theory, a "true score" on a test can be thought of as the average (hypothetical) score that would be obtained if an individual were to take the test an infinite number of times. In other words, the true score is the remaining part of the score once it has been stripped of random error.

In testing, one goal is to determine true score difference. What we know is:

$$X = T + e$$

where: X represents an observed score,
T represents a true score, and
e represents an error score (the part of the score due to random, nonsystematic influences on the test).

We only know an individual's observed score (X), which is their measured score on a test. With one equation and two unknowns, we cannot directly compute their true score, which is what we really want to know. Stated another way, we only know a person's measured or observed score; we want to know their true score. But, by definition, we cannot ever know a true score. So we must somehow estimate error and then estimate a true score.

To estimate true scores, we will work with the variances of observed scores, true scores, and error scores with the resulting formula:

$$\sigma_X^2 = \sigma_T^2 + \sigma_e^2$$

This formula indicates that the total variance in a distribution of observed scores is equal to the sum of the true variance plus the error variance.

Test reliability then gets defined as the ratio of true score variance (actual differences) to observed score variance (obtained differences); in other words, the proportion of variance in the observed scores that can be accounted for by variance in true scores.

The reliability of a test is designated r_{11} to indicate that the same construct is being measured twice with the same instrument:

$$r_{11} = \frac{\sigma_T^2}{\sigma_X^2}$$

If there were no error in the scores, and all of the scores were equal to the true scores, then the calculated r_{11} would be equal to 1. If all of the observed scores in a distribution contained equal parts of error and true ability (or whatever is being measured), then the value of r_{11} would be .5. In other words, if all the variance is accounted for by true variance, then r_{11} would be 1.00; if none of it is, then the value of r_{11} would be 0.

If r_{11} is the proportion of the observed score variance that is accounted for by true score variance, then $1 - r_{11}$ is the proportion of observed score variance accounted for by error. We can then rewrite the equation as follows:

$$1 - r_{11} = \frac{\sigma_e^2}{\sigma_X^2}$$

Since we will never know the true score variance, we are left with estimating a test's reliability. If we can obtain an estimate of the variance in the error, then we can estimate a person's true score.

The error variance can be estimated with the following formula:

$$\sigma_e^2 = \sigma_X^2 (1 - r_{11})$$

We have just explained Classical test theory. Now we'll see how we obtain reliability estimates to evaluate a test.

The reliability of a test is expressed as a positive decimal number ranging from 0.00 to 1.00: $r_{11} = 1.00$ indicates perfect reliability, and $r_{11} = 0.00$ indicates the complete absence of reliability. Reliability is usually estimated by analyzing the effects of variations in administration conditions (test-retest) and test content (parallel or alternate forms) on test taker's scores. Reliability is not affected by systematic changes in scores that affect all test takers similarly, but only unsystematic changes that have different

effects on different test takers. Such unsystematic factors influence the error variance of the test and hence the test's reliability. Reliability is estimated by calculating a correlation coefficient such as the Pearson r or Spearman's rho (depending on which is appropriate for the data).

At this point, we can use the Pearson r correlation coefficient to determine test-retest and interscorer reliability for a set of data. Consider the following situation: The Valdosta School of the Arts is accepting applications for the Fall semester. Recently, people have argued that past admissions procedures may no longer be appropriate. In the past, two criteria were examined 1) the score on a creative potential exam and 2) a judge's rating of an interpretive performance. Critics of the process have charged that scores on creative potential exams are not reliable and can change dramatically from one administration to the next; further, judges ratings are not reliable and vary from judge to judge. In an effort to determine the integrity of their admissions procedure, the school has decided to give the same exam twice to a large number of applicants and determine the statistical reliability using the Pearson r . Additionally, the reliability of the judges ratings will be examined by having a second judge rate the applicants' performance and then calculate the reliability of the interscorer rating. A subset of the data are shown here:

Student	Exam score	Retest score	Student	Judge 1 rating	Judge 2 rating
Kandi	38	37	Kandi	6	8
Steven	44	43	Steven	9	8
Andrew	28	30	Andrew	5	4
Cathy	35	36	Cathy	6	6
Craig	40	38	Craig	8	7
Leigh	47	45	Leigh	7	8
Pam	33	40	Pam	7	6
Jackie	30	28	Jackie	5	7
Veronica	34	35	Veronica	3	4
Shelli	48	46	Shelli	6	7

Following is the raw score formula for calculating the Pearson r :

$$r = \frac{N \sum XY - (\sum X \sum Y)}{\sqrt{[N \sum X^2 - (\sum X)^2][N \sum Y^2 - (\sum Y)^2]}}$$

Apply this formula to the data above letting X equal the first score on the exam and Y equal the score received on retest. Similarly for the interscorer data, X can equal the first judge's rating and Y the second judge's rating.

Step by step instructions for calculating the Pearson r follow:

Step 1

Setup a table with the 5 headings below listed across the top:

X , X^2 , Y , Y^2 , and XY

Step 2

List the values for X next to the corresponding values for Y in the table where they belong. Add all values for X , placing that sum at the bottom of the X column. Add all values for Y and place that sum at the bottom of the Y column.

Step 3

Multiply each value of X by its corresponding Y value and place that product in the corresponding XY column. Add all the XY products and place that sum at the bottom of the XY column.

Step 4

Square all the values of X and add all these values, placing the sum at the bottom of the X^2 column. Square all the values of Y and add these values together, placing the sum at the bottom of the Y^2 column.

Step 5

Insert the numbers into the formula and solve for r .

YOUR HOMEWORK:

In this assignment, you will practice calculating and interpreting reliability estimates. Use Pearson's r for the correlation coefficient. You are asked to make a decision about the admissions procedures based on your results.

Name _____ Section _____ Date _____

RELIABILITY- WORKSHEET

Homework

- 1) Calculate the correlation coefficient (r) between the first and second administration of the test given to the applicants. This constitutes test-retest reliability. This assignment can be complete by computer provided you show all the required information (see below).

X	X ²	Y	Y ²	XY
---	----------------	---	----------------	----

- 2) Calculate the correlation coefficient (r) between the judges' ratings. This constitutes inter-rater reliability.

X	X ²	Y	Y ²	XY
---	----------------	---	----------------	----

