Interpreting Correlation, Reliability, and Validity Coefficients: A Mix of Theory and Standards

---

Interpreting a Correlation Coefficient

I. Interpretation: Involves exploring the magnitude, direction, and probability

II. Theory: How do we expect two variables to be related to one another?

III. Magnitude:

- A. +/- .10 to +/- .39: low
- B. +/- .40 to +/- .69: moderate
- C. +/- .70 to +/- 1.00: high

IV. Direction

- A. + : variables increase or decrease in the same direction;
- B. -: variables increase or decrease in the opposite direction

V. Probability

- A. ..p < .05 (generally accepted as "criterion" of a significant/meaningful correlation)
  1. Has to do with the size of the sample, and the estimation of the true relationship between variables
  2. "There is a less than 5% chance that the relationship we are observing is due to chance alone." OR "We can be 95% certain that our relationship is a true one."
  3. May also be p < .01 or p < .001 (less than 1%; less than .1%)
- B. The most important criteria, because it tells us of it's meaningfulness and error

---

The Reliability Coefficient

I. Theoretically: Interpretation is dependant upon how stable we expect the construct we are measuring to be; likely, will vary with time

- A. *In decreasing order*, we would expect reliability to be highest for:
  1. Internal Consistency (Inter-Item): because all of our items should be *assessing the same construct*
  2. Alternate Forms: because, even though the items are different, they should *all be assessing the same construct*
  3. Test-Retest: most highly volatile because our construct may, theoretically, change over time
     - (i) Key question here becomes, "Is the consistency higher than what we would expect just due to chance alone?"
     - (ii) Additional question would be, "How much is due to chance error, and how much is due to true variability?"
       - (a) Coefficient of Determination Helps Here (higher = better)
       - (b) As does the Standard Error of Measurement
     - (iii) Final question involves, "For what purpose are we using the test?" The greater the potential impact of a test result, the greater we want this number to be.
     - (iv) The more stable we expect a construct to be, the higher the correlation we would want (e.g., magnitude)
     - (v) The further the time between Time 1 & Time 2, the lower we might expect the correlation to be
- B. Inter-rater Reliability: falls somewhere between alternate forms and test-retest; our raters are assessing the same construct, so we would expect or want this to be high
  1. However, again dependent upon how "easy" the construct is to observe and rate
- C. We would always expect this to be statistically significant.

II. Interpretation

- A. Same as with correlation coefficients, but interpreted with practical and theoretical considerations in mind.

The Validity Coefficient

I. Theoretical: We are attempting to see how our test (a) is able to predict constructs it should, theoretically, be able to predict, both over time and concurrently, and (b) compares to similar and dissimilar measures of the same or similar and different or dissimilar constructs

A. In both cases, we would expect lower correlations (than in reliability), because *we are comparing our test to different/other tests* (even when they are measuring the same construct)

1. Our same standards of judging a correlation coefficient still stand, but because there are so many other variables (e.g., potential explanations) involved what we typically expect and obtain is lower (e.g., ".50 & up are excellent")
2. A big potential limitation here is the psychometric properties of the criterion we relate our test to

B. In case "a," we would expect (1) the relationship between more similar tests to be higher (e.g., IQ and GPA v. Interest in Computers and Job Satisfaction); (2) the relationship between tests to be higher the less time that has expired between giving our test and being measured on the criterion

C. In case "b," we would expect (1) higher correlations between like measures, but these should not be "perfect" correlations. Theoretically, we are developing a test because we think it measures a construct "differently, better, or more effectively" than those measures already out there. Thus, we would expect our test to measure something "unique" or "new." (thus, less than perfect correlations); (2) lower correlations between dissimilar measures. However, theoretically, this may not be "complete differences" all the time (e.g., love and hate), but instead may have some overlap (e.g., anxiety and depression; math and logical ability). The important thing here is that we are able to explain, theoretically, the similarities and differences

D. Again, with any of these, we need to consider the *coefficient of determination*. This is really important in deciding how we might use a test.

1. For example, if the correlation between IQ and GPA is r=.50, this means IQ can explain 25% of variability in GPA. Thus, if we are doing an assessment to determine how well a child might do in school, we would want to ask ourselves: "In addition to this 25% that I can explain, what else do I want to know?" To decide this, our best bet would be to turn to the literature and say, "What other variables out there have been shown to be related to academic success or failure?" We would then want to include those things in a comprehensive evaluation.