The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation

Standard Error of Measurement

Contributors: Dorothy J. Musselwhite & Brian C. Wesolowski
Edited by: Bruce B. Frey
Book Title: The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation
Chapter Title: "Standard Error of Measurement"
Pub. Date: 2018
Access Date: February 26, 2018
Publishing Company: SAGE Publications, Inc.
City: Thousand Oaks,
Print ISBN: 9781506326153
Online ISBN: 9781506326139
DOI: http://dx.doi.org/10.4135/9781506326139.n658
Print pages: 1588-1590

The term *standard error of measurement* indicates the spread of measurement errors when estimating an examinee's true score from the observed score. Standard error of measurement is most frequently useful in test reliability. An observed score is an examinee's obtained score, or raw score, on a particular test. A true score would be determined if this particular test was then given to a group of examinees 1,000 times, under identical conditions. The average of those observed scores would yield the best estimate of the examinees' true abilities. Standard deviation is applied to the average of those scores across persons and administrations to determine the standard error of measurement. Observed score and true score can be used together to determine the amount of error:

$$\text{Score}_{\text{true}} = \text{Score}_{\text{observed}} + \text{Score}_{\text{error}}.$$

However, this true score is purely hypothetical and is not a practical way to estimate error. Therefore, other estimates of error must be used, including standard deviation and reliability.

Standard error of measurement applies to a single score and should be applied more frequently than a reliability coefficient to interpret individual score meaning. Standard error is used in conjunction with the normal distribution in order to make decisions about individual test scores. Accordingly, standard error can be used to estimate a range of scores around a specified cut point when determining an examinee's ability or potential. The normal distribution can aid in the interpretation of scores that fall above, below, or between specific points on the distribution. This concept is particularly important, as it relates to standardized testing and promotion or retention criteria. For example, if the cut point for failing is a 50, and administrators want to be 68% sure of their decision, standard error of measurement indicates that examinees who are within one standard error (*SE*) of the cut point (i.e., 50 ± $SE_{\text{measurement}}$) may fluctuate above or below the cut point if the test were administered again. In situations such as this, it is imperative that more data be gathered, such as class performance indicators or growth scores, to determine promotion or retention.

A large standard error indicates a large amount of variability between different samples; therefore, the sample may not accurately represent the population. This occurs when sample means are spread far along the *y*-axis, in the tails of the normal distribution. When sample means are grouped closer to the population mean, standard error will be smaller.

This entry first discusses the distinction between standard error and standard deviation, as these concepts are often confused. Then, standard error is applied to confidence intervals and other assessment situations.

Standard Error Versus Standard Deviation

Standard deviation indicates how well the mean represents sample data. When considering a population, however, the mean of one sample does not necessarily represent the mean of every possible sample. If several samples were taken from one population, each sample mean may differ. Sampling variation is crucial to understanding the connection from standard deviation to standard error.

Standard deviation is a measure of spread, specifically as scores are situated around the mean. More specifically, standard deviation considers scores between examinees. Standard deviation is more closely related to range but is not as affected by outlying scores. A high standard deviation is an indication that scores have more variation or are widely distributed around the mean. A low standard deviation indicates the scores have less variation and are not widely distributed around the mean.

Standard deviation is most useful when determining where examinees are expected to fall within a range of scores. For example, as standard deviation aligns with the concept of normal distribution, it can be assumed that roughly 68% of examinees will attain scores within the range of one standard deviation above and below the mean score. Standard deviation can be calculated using just one test administration.

Standard error is, like standard deviation, a measure of spread. Standard error determines an individual examinee's spread had that student been tested repeatedly. Unlike standard deviation, standard error must be calculated using a much larger data set. Using one sample numerous times would elicit very similar means among the test administrations. However, numerous random samples would yield many different means. These sample means would form their own normal distribution where the means would ultimately have a single mean. The "mean of means" would be the best approximation of the population mean. The standard deviation of the distribution of these means is called the standard error of the mean, which refers to the fluctuations, or errors, that occur when estimating the population mean from sample means. The distribution created by the statistics from these multiple samples is called the sampling distribution. Because it is not feasible to take 1,000 random samples, a formula is used to estimate standard error of the mean using one sample:

$$SE_{\text{mean}} = s/\sqrt{N},$$

where SE_{mean} refers to the standard error of the mean, s refers to the standard deviation of the mean, and N refers to the sample size. In terms of sample size, there exists an inverse relationship. Larger sample sizes will yield a smaller standard error, and smaller sample sizes will yield a larger standard error.

Standard error of measurement is different from standard error of the mean. Standard error of measurement focuses more on the spread of errors, as they relate to a true score compared to an observed score. Standard error of the mean focuses on error in relation to the estimation of population mean. In total, standard error investigates how varying the sample statistic is when numerous samples are extracted from the same population.

The standard error of a sampling distribution is calculated using multiple samples from the population in question. However, multiple samples are not always available, especially when administering a single test administration. Obtaining an infinite number of test administrations would be difficult due to money and time constraints but would also yield unfavorable effects, such as testing fatigue. Therefore, the researcher must assume that each individual test score is the best estimate of that examinee's true score. Similar to the estimation of population mean, sampling errors in the estimation of true scores additionally occur. Similarly, these sampling errors will also be normally distributed, with a standard deviation called the standard error of measurement. The estimate for the standard error of measurement is calculated using the following formula:

$$SE_{\text{measurement}} = s\sqrt{1 - r_{xx}},$$

where $SE_{\text{measurement}}$ refers to the standard error of measurement, s refers to the standard deviation of the measure or test, and r_{XX} refers to the reliability of the measure (e.g., Cronbach's α).

The reliability coefficient represents the test's consistency. The aforementioned formula shows that the standard error of measurement increases as the standard deviation increases. In addition, the standard error of measurement increases as the test reliability decreases, showing an inverse relationship. If a test is perfectly reliable ($r = 1.0$), an examinee will attain the same score for every test administration. If the reliability is close to perfect, the standard error will be small, indicating the examinee's observed score is very similar to the true score.

Confidence Intervals

Standard error of measurement can be most beneficial in the construction of confidence intervals. Standard error and standard deviation are similar in that they both explore estimates of true scores. Under the assumptions of the normal distribution, 68% of the time, examinees' true scores lie within one standard error of measurement above or below the mean (±1). Next, 96% of the time, examinees' true scores would lie within two standard errors of measurement above or below the mean (±2). Last, 99.7% of the time, examinees' true scores would lie within three standard errors of measurement above or below the mean (±3).

The standard error is combined with the examinee's observed score, just as standard deviation is, to determine the upper and lower bound of the confidence interval for that specific examinee. Standard error can only apply to an individual score when developing a confidence interval. Standard error is often associated with probability and the prediction of true scores, which should be applied to the distribution of scores as a whole.

An example of standard error of measurement with confidence intervals can be illustrated through intelligence testing, such as the IQ test. On one test administration, a student may earn a score of 108. Other students in the same testing administration may have received similar scores, and a standard error of 5 was calculated. Using each student's score and the associated group standard error, an interval can be calculated to determine where each student is likely to score if they were given the test again. The standard error is added and subtracted to the original score ($108 + 5 = 113$, $108 - 5 = 103$). This student is 68% likely to score between 103 and 113 if given the IQ test again. The interval can be increased by adding and subtracting the standard error again, thereby becoming more likely to predict the student's score on the next administration. As the confidence level increases, precision will decrease, but Type I error rate will decrease.

Standard Error of Estimate

Using a regression analysis, the standard error of estimate approximates how spread the prediction errors are when using *X* values to predict *Y* values. These errors occur because of unreliable measurement in one of the variables or because of unsystematic differences between the values. The regression analysis provides a best estimate as to an examinee's predicted score, but similar to other standard error measures, there will be sampling errors around the estimate. The standard error of estimate should not be used as an estimator of true scores when comparing to observed scores.

Standard error of measurement is used to express test reliability. Standard error of estimate is used to express test validity. A small standard error of estimate indicates a more valid test.

Assessment and Measurement

Assessments are more frequently being used as a method of describing individuals, even as an indicator of the examinee's fate. The raw score gives little information as to an examinee's ability and characterization. The interpretation of these scores is becoming more essential in determining students' actual abilities. Standard error of measurement serves as an indicator of reliability that is independent of the variability in sample groups. Using the idea of confidence intervals, a student's ability can be perceived as a range of scores rather than one specific score. This concept of bands of scores can apply not only to one individual on different tests but also to multiple students on the same test.

See alsoNormal Distribution; Reliability; Standard Deviation; Standard Error of Measurement; Validity

Dorothy J. MusselwhiteBrian C. Wesolowski

http://dx.doi.org/10.4135/9781506326139.n658
10.4135/9781506326139.n658

Further Readings

Boyle, J. D., & Radocy, R. E. (1987). Measurement and evaluation of musical experiences. New York, NY: Schirmer Books.
Field, A. (2013). Discovering statistics using IBM SPSS statistics. London, UK: Sage.
Gravetter, F. J., & Wallnau, L. B. (2011). Essentials of statistics for the behavioral sciences. Belmont, CA: Wadsworth Publishing.
Kubiszyn, T., & Borich, G. (2003). Educational testing and measurement: Classroom application and practice. New York, NY: Wiley.
Payne, D. A. (2003). Applied educational assessment. Toronto, Canada: Wadsworth Publishing.
Reid, H. M. (2014). Introduction to statistics: Fundamental concepts and procedures of data analysis. Thousand Oaks, CA: Sage.