GAMALEARN
Experience. Innovate

SwiftAssess CTT (CLASSICAL TEST THEORY STATISTICS)

GamaLearn

Test and Item Statistics Documentation

GAMALEARN
Experience. Innovate

Table of Contents

General Statistics

Mode

In statistics, the mode is the value that is repeatedly occurring in a given set. In the case of grouped frequency distribution, calculation of mode just by looking into the frequency is not possible. To determine the mode of data in such cases we calculate the modal class. Mode lies inside the modal class. The mode of data is given by the formula:

l = lower limit of the modal class, h = size of the class interval

f_1 = frequency of the modal class

f_0 = frequency of the class preceding the modal class

f_2 = frequency of the class succeeding the modal class

$$Mode = l + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) \times h$$

Variance

Variance can simply be defined as a measure of variability to represent members of a group. The variance measures the closeness of data points corresponding to a greater value of variance. Variances describe the variability of the observed observations.

Population variance

When you have collected data from every member of the population that you're interested in, you can get an exact value for population variance.

The population variance formula looks like this:

Variance = $\frac{\sum(\mathbf{Xi} - \boldsymbol{\mu})^2}{\boldsymbol{N}}$, where **Xi** = i_{th} data point in the data set, **μ** = Population mean, **N** = Number of data points in the population

When you collect data from a sample, the sample variance is used to make estimates or inferences about the population variance.

The sample variance formula looks like this:

Variance = $\frac{\sum(\mathbf{Xi} - \boldsymbol{\mu})^2}{\boldsymbol{N-1}}$, where **Xi** = i_{th} data point in the data set, **μ** = Sample mean, **N** = Number of data points in the sample

With samples, we use n – 1 in the formula because using n would give us a biased estimate that consistently underestimates variability. The sample variance would tend to be lower than the real variance of the population.

Reducing the sample n to n – 1 makes the variance artificially large, giving you an unbiased estimate of variability: it is better to overestimate rather than underestimate variability in samples.

It's important to note that doing the same thing with the standard deviation formulas doesn't lead to completely unbiased estimates. Since a square root isn't a linear operation, like addition or subtraction, the unbiasedness of the sample variance formula doesn't carry over the sample standard deviation formula.

Standard Deviation

Standard deviation observes the quantifiable amount of dispersion of observations when approached with data. Standard deviation measures the dispersion of observations within a set. Standard Deviation = $\sqrt{Variance}$

Standard Error of Mean

It is the standard deviation of the sampling distribution of the mean. The formula for the standard error of the mean is: where σ is the standard deviation of the original distribution and N is the sample size (the number of scores each mean is based upon). This formula does not assume a normal distribution. However, many of the uses of the formula do assume a normal distribution. The formula shows that the larger the sample size, the smaller the standard error of the mean. More specifically, the size of the standard error of the mean is inversely proportional to the square root of the sample size.

$$\sigma_M = \frac{\sigma}{\sqrt{N}}$$

In R implementation, the '*plotrix*' library package calculates the conventional standard error of the mean = sd(x)/sqrt(sum((x)))

Test-based Statistics

Skewness

Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point. It tells us how observations are distributed around the mean.

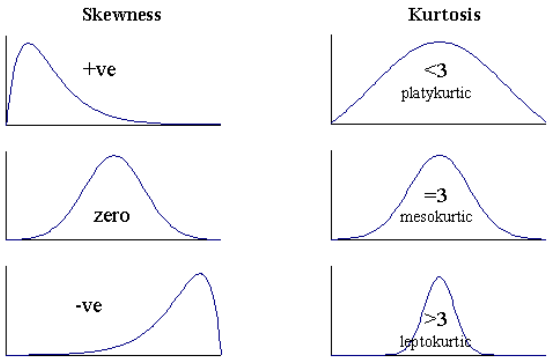
For univariate data Y1, Y2, ..., YN, the formula for skewness is:

$$g_1 = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^3/N}{s^3}$$

where $\bar{Y}$ is the mean, s is the standard deviation, and N is the number of data points. Note that in computing the skewness, the s is computed with N in the denominator rather than N - 1.

The above formula for skewness is referred to as the Fisher-Pearson coefficient of skewness. Many software programs actually compute the adjusted Fisher-Pearson coefficient of skewness.

In R implementation, the '*moments*' and 'e1071' library package calculates the Skewness using the Fisher-Pearson coefficient of skewness as stated above.



Kurtosis

Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. That is, data sets with high kurtosis tend to have heavy tails, or outliers. Data sets with low kurtosis tend to have light tails, or lack of outliers. A uniform distribution would be the extreme case.

For univariate data Y1, Y2, ..., YN, the formula for kurtosis is:

$$\text{kurtosis} = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^4/N}{s^4}$$

where $\bar{Y}$ is the mean, s is the standard deviation, and N is the number of data points. Note that in computing the kurtosis, the standard deviation is computed using N in the denominator rather than N - 1.

The kurtosis for a standard normal distribution is three. For this reason, some sources use the following definition of kurtosis (often referred to as "excess kurtosis"):

$$\text{kurtosis} = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^4/N}{s^4} - 3$$

This definition is used so that the standard normal distribution has a kurtosis of zero. In addition, with the second definition positive kurtosis indicates a "heavy-tailed" distribution and negative kurtosis indicates a "light tailed" distribution.

In R implementation, the 'moments' library package calculates the Kurtosis using the Pearson's measure of kurtosis.

Pearson's moment coefficient of kurtosis (excess kurtosis)

$$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3$$

Variables

- γ_2 Excess kurtosis (dimensionless)
- μ_4 The fourth moment around the mean (dimensionless)
- σ The standard deviation (dimensionless)

Test Reliability (Cronbach's Alpha)

Reliability indicates how reliably or precisely a questionnaire or test measures a true value. Reliability therefore means how accurately a test can measure a variable. The reliability of a test is higher the fewer measurement errors there are.

Cronbach's alpha is a measure of the internal consistency of a scale. Cronbach's alpha is thus a measure of the extent to which the group of questions are related to one another and thus provides an estimate of how good or poor the measurement accuracy, known as reliability, of a group of items is.

Suppose that we measure a quantity which is a sum of K components (*K-items* or *testlets*): $X = Y_1 + Y_2 + \cdots + Y_K$. Cronbach's α is defined as

$$\alpha = \frac{K}{K-1}\left(1 - \frac{\sum_{i=1}^{K} \sigma_{Y_i}^2}{\sigma_X^2}\right)$$

where σ_X^2 is the variance of the observed total test scores, and $\sigma_{Y_i}^2$ the variance of component *i* for the current sample of persons.[4]

The Cronbach's alpha thus becomes larger when the number of items is increased and when the inter-item correlation increases. The Cronbach's alpha becomes smaller when the average inter-item correlation becomes smaller.

Cronbach's alpha	Internal consistency
α ≥ 0.9	Excellent
0.9 > α ≥ 0.8	Good
0.8 > α ≥ 0.7	Acceptable
0.7 > α ≥ 0.6	Questionable
0.6 > α ≥ 0.5	Poor
0.5 > α	Unacceptable

In R implementation, the '*ltm*' package uses this formula to calculate the Cronbach Alpha:

The Cronbach's alpha computed by `cronbach.alpha()` is defined as follows

$$\alpha = \frac{p}{p-1}\left(1 - \frac{\sum_{i=1}^{p} \sigma_{y_i}^2}{\sigma_x^2}\right),$$

where p is the number of items σ_x^2 is the variance of the observed total test scores, and $\sigma_{y_i}^2$ is the variance of the ith item.

Standard Error of Measurement

In classical test theory, it is defined as **SEM = SD * sqrt (1 – r)**

Where SD is the standard deviation of scores for everyone who took the test, and r is the reliability of the test. It is interpreted as the standard deviation of scores that you would find if you had the person take the test over and over, with a fresh mind each time. A confidence interval with this is then interpreted as the band where you would expect the person's true score on the test to fall.

Item-based Statistics

Item Difficulty P-Value

The item difficulty index is a common and particularly useful analytical tool for statistical analysis, especially when it comes to determining the validity of test questions in an educational setting. The item difficulty index is often called the p-value because it is a measure of proportion – for example, the proportion of students who answer a particular question correctly on a test. P-values are found by using the difficulty index formula, and they are reported in a range between 0.0 and 1.0. In the scenario with students answering questions on a test, higher p-values, or p-values closer to 1.0, correspond with a greater proportion of students answering that question correctly. In other words, easier test questions will have greater p-values. That is why some statisticians also call the difficulty index "the easiness index" when they are performing an item analysis on data sets that have to do with education.

The formula looks like this: the number of students who answer a question correctly (c) divided by the total number of students in the class who answered the question (s). **The formula: c ÷ s = p**

The answer will equal a value between 0.0 and 1.0, with harder questions resulting in values closer to 0.0 and easier questions resulting in values closer to 1.0.

Item Total Correlation Discrimination

The discrimination index is another way that test writers can evaluate the validity of their tests. Item discrimination evaluates how well an individual question sorts students who have mastered the material from students who have not. Test takers with understanding of the material should be more likely to answer a question correctly, whereas students without knowledge of the material should get the question wrong. Questions that do an excellent job of sorting those students who have mastered the material from students who have not are called "highly discriminating."

The item discrimination index is a measure of how well an item is able to distinguish between examinees who are knowledgeable and those who are not, or between masters and non-masters. There are actually several ways to compute an item discrimination, but one of the most common is the point-biserial correlation (r_{pb}). This statistic looks at the relationship between an examinee's performance on the given item (correct or incorrect) and the examinee's score on the overall test. For an item that is highly discriminating, in general the examinees who responded to the item correctly also did well on the test, while in general the examinees who responded to the item incorrectly also tended to do poorly on the overall test.

$$r_{pb} = \frac{(\bar{y}_1 - \bar{y}_2).\sqrt{pq}}{s_y}$$

where, p = proportion for which nominal value is 1,
q = proportion for which nominal value is 0,
$\bar{y}_1$ = conditional mean of the quantitative or numerical variable y when the nominal score is 1,
$\bar{y}_2$ = conditional mean of the quantitative or numerical variable y when the nominal score is 0, and
s_y = standard deviation of the numerical feature.

The possible range of the discrimination index is -1.0 to 1.0; however, if an item has a discrimination below 0.0, it suggests a problem. When an item is discriminating negatively, overall, the most knowledgeable examinees are getting the item wrong, and the least knowledgeable examinees are getting the item right. A negative discrimination index may indicate that the item

is measuring something other than what the rest of the test is measuring. More often, it is a sign that the item has been mis-keyed.

When interpreting the value of a discrimination it is important to be aware that there is a relationship between an item's difficulty index and its discrimination index. If an item has an extremely high (or very low) p-value, the potential value of the discrimination index will be much less than if the item has a mid-range p-value. In other words, if an item is either very easy or very hard, it is not likely to be very discriminating. A useful approach when reviewing a set of item discrimination indexes is to also view each item's p-value at the same time. For example, if a given item has a discrimination index below .1, but the item's p-value is greater than .9, you may interpret the item as being easy for the complete set of examinees, and for that reason not providing much discrimination between high ability and low ability examinees.

High-Low Discrimination

The interpretation of High-Low Discrimination is similar to the interpretation of correlational indices: positive values indicate good discrimination, values near zero indicate that there is little discrimination, and negative discrimination indicates that the item is easier for low-scoring participants.

To calculate the High-Low Discrimination value, we simply subtract the percentage of low-scoring participants who got the item correct from the percentage of high-scoring participants who got the item correct. If 30% of our low-scoring participants answered correctly, and 80% of our high-scoring participants answered correctly, then the High-Low Discrimination is 0.80 – 0.30 = 0.50.

In Measuring Educational Achievement, Ebel recommended the following cut points for interpreting High-Low Discrimination (D):

D Range	Interpretation
0.40 ≤ *D* ≤ 1.00	Satisfactory discrimination
0.30 ≤ *D* < 0.40	Some revisions may be required to the item
0.20 ≤ *D* < 0.30	The item needs revision
-1.00 ≤ *D* < 0.20	The item needs to be removed or completely revised

Item Reliability

The item-reliability index provides an indication of the internal consistency of a test, the higher this index, the greater the test's internal consistency. Item reliability is simply the product of the standard deviation of item scores and a correlational discrimination index (Item-Total Correlation Discrimination in the Item Analysis Report). So, item reliability reflects how much the item is contributing to total score variance. As with assessment reliability, higher values represent better reliability.

This index is equal to the product of the item-score standard deviation (s) and the correlation (r) between the item score and the total test score.

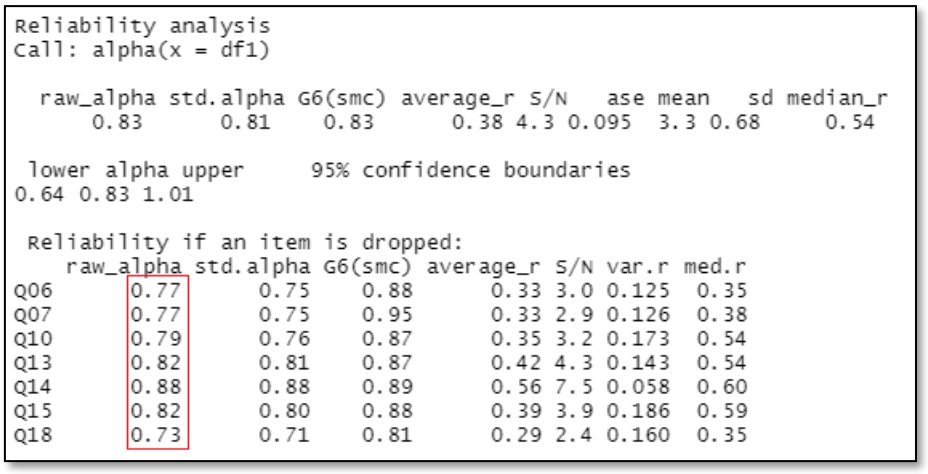
Alternative Method

R's Technique

Item reliability is the consistency of a set of items (variables); that is to what extent they measure the same thing. When a set of items are consistent, they can make a measurement scale such as a sum scale. Cronbach's alpha is the most popular measure of item reliability; it is the average correlation of items in a measurement scale. If the items have variances that significantly differ, standardized alpha is preferred.

When all items are consistent and measure the same thing, then the coefficient alpha is equal to 1. A high value for alpha does not imply that the measure is unidimensional. To prove that a scale is unidimensional, you can use factor analysis to check the dimensionality.

It is possible to see the effect of an individual item on the overall alpha value by recomputing Cronbach's alpha excluding that item. This can be seen in the figure below where the values captured in the red box signify effect of an individual item on the overall alpha by eliminating the item in the Cronbach Alpha Calculation. If alpha increases when you exclude an item, that item does not highly correlate with the other items in the scale. If the alpha decreases, that item does correlate with the other items in the scale.



```
Reliability analysis
Call: alpha(x = df1)

  raw_alpha std.alpha G6(smc) average_r S/N   ase mean   sd median_r
      0.83      0.81    0.83      0.38 4.3 0.095  3.3 0.68     0.54

 lower alpha upper     95% confidence boundaries
0.64 0.83 1.01

 Reliability if an item is dropped:
    raw_alpha std.alpha G6(smc) average_r S/N var.r med.r
Q06      0.77      0.75    0.88      0.33 3.0 0.125  0.35
Q07      0.77      0.75    0.95      0.33 2.9 0.126  0.38
Q10      0.79      0.76    0.87      0.35 3.2 0.173  0.54
Q13      0.82      0.81    0.87      0.42 4.3 0.143  0.54
Q14      0.88      0.88    0.89      0.56 7.5 0.058  0.60
Q15      0.82      0.80    0.88      0.39 3.9 0.186  0.59
Q18      0.73      0.71    0.81      0.29 2.4 0.160  0.35
```

References

Mode

Mode | Mode in Statistics (Definition, How to Find Mode, Examples) (byjus.com)

Variance and Standard Deviation:

Difference Between Variance and Standard Deviation | Comparison (byjus.com)

What is Variance? | Definition, Examples & Formulas (scribbr.com)

Skewness and Kurtosis

1.3.5.11. Measures of Skewness and Kurtosis (nist.gov)

Skewness & Kurtosis Simplified. What is Skewness and how do we detect… | by Atul Sharma | Towards Data Science

Skewness and Kurtosis |Shape of data: Skewness and Kurtosis (analyticsvidhya.com)

Item Reliability

Item Analysis: The Item-Reliability Index (lahoktraining.blogspot.com)

Item reliability > Statistical Reference Guide | Analyse-it® 5.65 documentation