In this handout, we will discuss different types of and methods for establishing *reliability*. Recall that this concept was defined in the previous handout as follows.

|  |
| --- |
| Definition |
| **Reliability** is the extent to which repeatedly measuring the same thing produces the same result. |

In order for survey results to be useful, the survey must demonstrate **reliability**. The best practices for questionnaire design discussed in the previous handout help to maximize the instrument’s reliability.   
 **THEORY OF RELIABILITY**

Reliability can be thought of as follows: .  
  
In some sense, this is the proportion of “truth” in a measure. For example, if the reliability is estimated to be .5, then about half of the variance of the observed score is attributable to truth; the other half is attributable to error. What do you suppose is the desired value for this quantity?  
  
Note that the denominator of the equation given above can be easily computed. The numerator, however, is unknown. Therefore, we can never really compute reliability; we can, however, *estimate* it. In the remainder of this handout, we will introduce various types of reliability relevant to survey studies and discuss how reliability is estimated in each case.

**TYPES AND MEASURES OF RELIABILITY RELEVANT TO SURVEY STUDIES**

When designing survey questionnaires, researchers may consider one or more of the following classes of reliability.

|  |
| --- |
| Types of Reliability |
| **Test-Retest Reliability** –this is used to establish the consistency of a measure from one time to another. |
| **Parallel Forms Reliability** – this is used to assess whether two forms of a questionnaire are equivalent. |
| **Internal Consistency Reliability** - this is used to assess the consistency of results across items within a single survey instrument. |

Each of these is discussed in more detail below.

**Test-Retest Reliability**

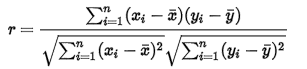
We estimate test-retest reliability when we administer the same questionnaire (or test) to the same set of subjects on two different occasions. Note that this approach assumes there is no substantial change in what is being measured between the two occasions. To maximize the chance that what is being measured is not changing, one shouldn’t let too much time pass between the test and the retest.  
  
There are several different measures available for estimating test-retest reliability. In particular, we will discuss the following in this handout:

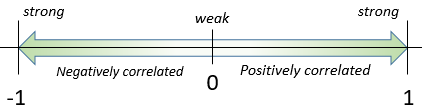
* Pearson’s correlation coefficient
* ICC (intraclass correlation coefficient)
* Kappa statistic

**Example 4.1**: Suppose we administer a language proficiency test and retest to a random sample of 18 students. Their scores from both time periods are shown below.

|  |  |
| --- | --- |
| Test – Retest  Data | library(ggplot2)  ggplot(TestData, aes(Test,Retest)) + geom\_point() + geom\_abline(intercept=0,slope=1) |

One way to assess test-retest reliability is to compute **Pearson’s correlation coefficient** between the two sets of scores. If the test is reliable and if none of the subjects have changed from Time 1 to Time 2 with regard to what is being measured, we should see a high correlation coefficient.





|  |  |
| --- | --- |
| *Strongest Reliability Possible* | |
|  | > cor(TestData[,2:3])  Test Retest  Test 1.0 1.0  Retest 1.0 1.0 |

Getting Pearson’s Correlation coefficient in R can be done as follows.

> library(dplyr)

> summarize(TestData,cor(Test,Retest))

cor(Test, Retest)

1. 0.7308101

OR

> cor(TestData[,2:3])

Test Retest

Test 1.0000000 0.7308101

Retest 0.7308101 1.0000000

Questions:

1. What is the Pearson correlation coefficient for the example given above?
2. Does this indicate that this test is “reliable”? Explain.

In addition to computing the correlation coefficient, one should also compute the mean and standard deviation of the scores at each time period.

summarize(TestData,"Avg Test"=mean(Test),"Avg Retest" = mean(Retest),"Std Dev Test"= sd(Test), "Std Dev Retest" = sd(Retest),"Num Subjects" = n())

Avg Test Avg Retest Std Dev Test Std Dev Retest Num Subjects

1 81.55556 83.61111 9.166578 10.65026 18

The Pearson correlation coefficient is an acceptable measure of reliability, but it has been argued that a better measure of test-retest reliability for continuous data is the **intraclass correlation coefficient (ICC)**. One reason the ICC is preferred is that Pearson’s correlation coefficient has been shown to overestimate reliability for small sample sizes. Another advantage the ICC has is that it can be calculated even when you administer the test at more than two time periods.   
  
There are several versions of the ICC, but one that is typically used in examples such as this is computed as follows:



where k = the number of time periods, MSSubject = the between-subjects mean square, and   
MSError = the mean square due to error after fitting a repeated measures ANOVA.  
  
Let’s compute the ICC for the data in Example 4.1.

|  |  |
| --- | --- |
| Stacked data format | First, Subject must be identified as a factor when modeling. Next, the aov() function can be used to obtain the necessary means squares for the calculations.  > str(TestDataICC)  'data.frame': 36 obs. of 3 variables:  $ Subject: int 1 2 3 4 5 6 7 8 9 10 ...  $ Time : Factor w/ 2 levels "Initial","Retest": 1 1 1 1  $ Score : int 83 78 63 96 83 76 62 86 86 76 ...  > summary(aov(Score ~ as.factor(Subject) + Time, data=TestDataICC))  Df Sum Sq Mean Sq F value Pr(>F)  as.factor(Subject) 17 2891.3 170.07 6.211 0.000237 \*\*\*  Time 1 38.0 38.03 1.389 0.254832  Residuals 17 465.5 27.38  Computing the intra-class correlation using the results from the ANOVA table above.  = |

Getting Intra-Class Correlation with the irr package in R.

|  |  |
| --- | --- |
| > library(irr) > icc(TestData[,2:3],model="oneway",type="agreement")  Single Score Intraclass Correlation  Model: oneway  Type : agreement  Subjects = 18  Raters = 2  ICC(1) = 0.718  F-Test, H0: r0 = 0 ; H1: r0 > 0  F(17,18) = 6.08 , p = 0.000201 |  |

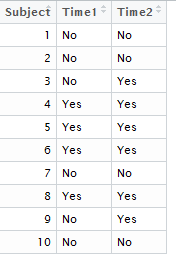
Getting Intra-Class Correlation with the psych package [Procedures for Psychological, Psychometric, and Personality Research] in R. See Shrout and Fleiss (1979) for discussion of various types of intra-class correlation that are computed here.

|  |
| --- |
| > library(psych)  > ICC(TestData[,2:3]) |

The exact wordage on interpretation of ICC is some fluid – reference to the Wiki page on intra-class correlation is provided here.

|  |  |
| --- | --- |
| Wiki Entry: <https://en.wikipedia.org/wiki/Intraclass_correlation> | |
|  |  |

In the previous example, the data were considered on a continuous scale. Note that when the data are measured on a binary scale, Cohen’s **kappa statistic** should be used to estimate test-retest reliability; for nominal data with more than two categories, one can use Fleiss’s kappa statistic. Finally, when the data are ordinal, one should use the **weighted kappa**.  
**Example 4.2**: Suppose 10 nursing students are asked on two different occasions if they plan to work with older adults when they graduate.



Cohen’s kappa statistic is a function of the number of agreements observed minus the number of agreements we expect by chance.

κ = 

Cohen’s kappa statistic is computed by first organizing the data in the 2x2 table as follows.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Time 2: No | Time 2: Yes | Total |
| Time 1: No | **4** | 2 | 6 |
| Time 1: Yes | 0 | **4** | 4 |
| Total | 4 | 6 | 10 |

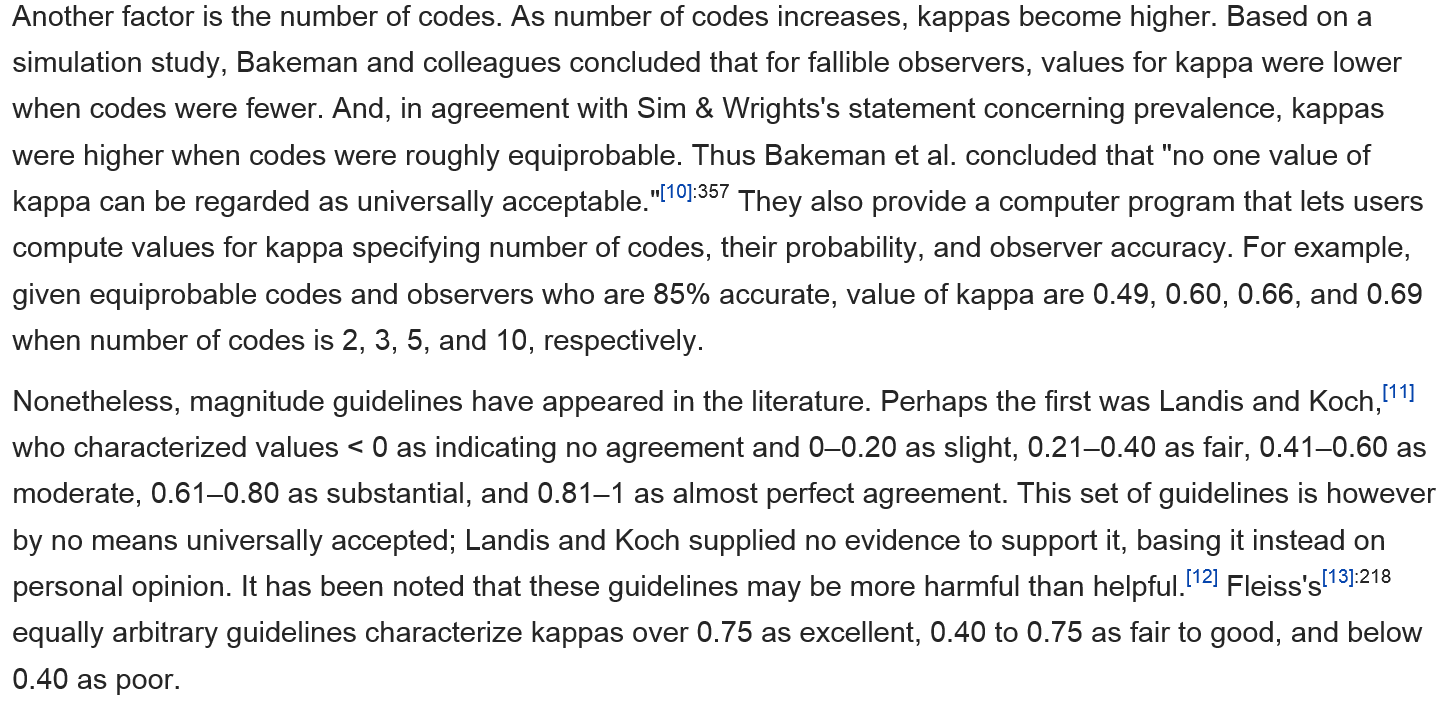
Computation for computing the number of expected by chance.  
*Note: These are simple the expected cell counts from a Chi-Square Test*

|  |  |  |  |
| --- | --- | --- | --- |
|  | Time 2: No | Time 2: Yes | Total |
| Time 1: No |  |  | 6 |
| Time 1: Yes |  |  | 4 |
| Total | 4 | 6 | 10 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Yes | No | Total |
| Agreements Observed |  |  |  |
| Agreements Expected by Chance |  |  |  |

κ = 

To interpret the kappa coefficient, consider the following discussion from wikipedia.org.



Question: Based on the kappa statistic obtained in Example 4.2, what can you say about the test-retest reliability of this question asked to nursing students?

Calculating the kappa statistic in R  
  
The data table is first arranged as follows:  
Getting Pearson’s Correlation coefficient in R can be done as follows.

> table(TestDataKappa[,2:3])

Time2

Time1 No Yes

No 4 2

Yes 0 4

Computing Cohen’s Kappa using the irr package in R.

> library(irr)

> kappa2(TestDataKappa[,2:3])

Cohen's Kappa for 2 Raters (Weights: unweighted)

Subjects = 10

Raters = 2

Kappa = 0.615

z = 2.11

p-value = 0.035

**Weighted Kappa**  
  
As mentioned earlier, if the data are ordinal, e.g. likert-scale, the **weighted kappa** should be used. This is a generalization of the kappa statistic that uses weights to quantify the relative difference between categories (essentially, disagreements that are further away from one another on the scale are weighted differently than disagreements that are closer together on the scale).

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 5 point Likert-scale  Image result for 5 point likert scale | |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | |  | | Retest | | | | | | SD | D | N | A | SA | | Test | SD |  |  |  |  |  | | D |  |  |  |  |  | | N |  |  |  |  |  | | A |  |  |  |  |  | | SA |  |  |  |  |  | |

Consider the following data from a test / re-test situation. Notice many of the combinations were not selected on either the Test or Re-test; thus, the data will be reduced to a 3x3 matrix.

|  |  |
| --- | --- |
| Original Data | Relevant data |

The idea of the weighted kappa statistic is to provide more weight to combinations for which agreement is closer. A commonly used weighting scheme is to use linear or absolute weighting, i.e. **Cicchetti-Allison** weights. The weights for each cell are computed as follows.

|  |  |
| --- | --- |
| Mathematical Formula  [Cicchetti-Allison Weights] | Weighting matrix |

Getting Weighted Kappa Statistic using the cohen.kappa() function from the psych package in R.

> library(psych)

> agreementdata <- matrix(c(6,3,1,1,15,6,1,5,22),ncol=3,byrow=TRUE)

> abs\_weights <- matrix(c(1.0,0.5,0.0,0.5,1.0,0.5,0.0,0.5,1.0),ncol=3,byrow=TRUE)

> cohen.kappa(agreementdata,w=linear\_weights)

Call: cohen.kappa1(x = x, w = w, n.obs = n.obs, alpha = alpha, levels = levels)

Cohen Kappa and Weighted Kappa correlation coefficients and confidence boundaries

lower estimate upper

unweighted kappa 0.35 0.54 0.72

weighted kappa 0.40 0.58 0.76

Number of subjects = 60

An alternative weighting scheme that is based on squared distance from diagonal.

|  |  |
| --- | --- |
| Mathematical Formula  [Fleiss-Cohen Weights] | Weighting matrix |

> agreementdata <- matrix(c(6,3,1,1,15,6,1,5,22),ncol=3,byrow=TRUE)

> sq\_weights <- matrix(c(1.0,0.75,0.0,0.75,1.0,0.75,0.0,0.75,1.0),ncol=3,byrow=TRUE)

> cohen.kappa(agreementdata,w=sq\_weights)

Call: cohen.kappa1(x = x, w = w, n.obs = n.obs, alpha = alpha, levels = levels)

Cohen Kappa and Weighted Kappa correlation coefficients and confidence boundaries

lower estimate upper

unweighted kappa 0.35 0.54 0.72

weighted kappa 0.45 0.63 0.81

Number of subjects = 60

**Parallel Forms Reliability**

This involves creating a large set of questions that are believed to measure the same construct and then randomly dividing these questions into two sets (known as parallel forms). Both sets are then administered to the same group of people. The means, standard deviations, and correlations with other measures (when appropriate) should be compared to establish that the two forms are equivalent. The correlation between the two parallel forms can be used as the estimate of reliability (we want the scores on the two forms to be highly correlated).  
  
Having parallel forms is useful for studies involving both a pre-test and post-test in which the researcher doesn’t want to use the same form at both time periods (think of why they may want to avoid this). If parallel forms reliability can be established, then the researcher can use both forms in their study.  
  
Question: If the researcher uses parallel forms, should they always use Form A for the pre-test and Form B for the post-test? Can you think of a different approach that might be better?

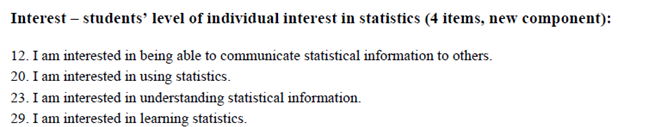
**Internal Consistency Reliability**Earlier, we discussed methods for establishing test-retest reliability. Note that in some cases, it may be too expensive or time-consuming to administer a test twice. Or, one may not want “practice effects” to influence the results which is always a possibility when a measurement instrument is administered more than once.   
  
In such cases, one is better off investigating **internal consistency reliability**. This involves administering a single measurement instrument (e.g., a survey questionnaire) to a group of people on only one occasion. Essentially, we judge the reliability of the instrument by estimating how well the items that measure the same construct yield similar results.

There are several different measures available for estimating internal consistency reliability. In particular, we will discuss the following in this handout:

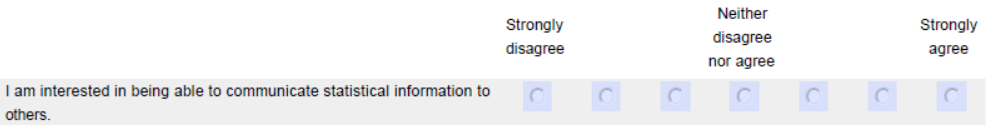
* Average inter-item correlation
* Split-half reliability
* Cronbach’s alpha

Sometimes, these measures are computed based on all items measured by the instrument; other times, these are used to establish the reliability associated with various constructs that are measured by different items within the same instrument. Next, we will introduce these measures of reliability in the context of an example.

**Example 4.3**. The Survey of Attitudes toward Statistics (SATS) measures six different components of students’ attitudes about statistics. The survey overall consists of 36 questions. One of the constructs that the survey measures is students’ “interest” in statistics, which is measured with the following four questions.



All of the questions are measured on a 7-point scale, as shown below for only one question.



To score this survey, the data is first recoded so that 1 = strongly disagree, 4 = neither disagree nor agree, and 7 = strongly agree, etc. Then, the score for the “interest” component is computed by averaging the responses of these four questions. Other components are scored in the same way.



Calculating the Average Inter-Item Correlation  
  
First, we will consider establishing reliability for only the “interest” scale using the average inter-item correlation to measure reliability.

> library(dplyr)

> SATS\_Interest <- select(SATSData,I12,I20,I23,I29)

> head(SATS\_Interest)

I12 I20 I23 I29

1 5 4 4 4

2 NA NA NA NA

3 5 5 5 6

4 5 7 6 6

5 5 5 5 5

6 NA NA NA NA

Getting the correlations in R

> cor(SATS\_Interest)

I12 I20 I23 I29

I12 1 NA NA NA

I20 NA 1 NA NA

I23 NA NA 1 NA

I29 NA NA NA 1

Getting the correlations in R and excluding the missingness

> cor( na.omit(SATS\_Interest) )

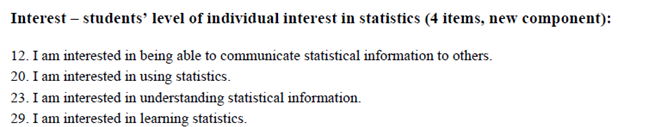
I12 I20 I23 I29

I12 1.0000000 0.5379079 0.5230786 0.4723748

I20 0.5379079 1.0000000 0.6844636 0.7344580

I23 0.5230786 0.6844636 1.0000000 0.7185934

I29 0.4723748 0.7344580 0.7185934 1.0000000



Questions:

1. Which of the Interest items have the highest internal consistency? The least?
2. Would you suggest more work be done to make these survey items more consistent? Explain.

Finally, computing the average correlations across these four interest items can be done as follows.

> cor\_output2 <- as.matrix(cor\_output)

> cor\_vector <- cor\_output[upper.tri(cor\_output2)]

> mean(cor\_vector)

[1] 0.6118127

Question:

1. Note that the average inter-item correlation is .612, with the individual correlations ranging from .47 to .72. What does this imply about the reliability of the “interest” construct?

Calculating the Split-Half Reliability  
To compute this measure of reliability, we correlate scores on one random half of the items on a survey (or test) with the scores on the other random half. Consider the calculation of this measure using the SATS data from Example 4.3.

> SATS\_Interest\_NoMissing <- na.omit(SATS\_Interest)

> head(SATS\_Interest\_NoMissing)

I12 I20 I23 I29

1 5 4 4 4

3 5 5 5 6

4 5 7 6 6

5 5 5 5 5

7 4 3 5 4

8 6 5 7 7

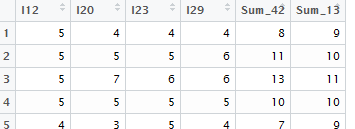
The following is used to get a random split of the columns. Here the average of column 4 and 2 will be computed, the average of columns 1 and 3 will be computed.

> sample(1:4,replace=F)

[1] 4 2 1 3

The total score for each split for each row will be computed using the mutate() function from the dplyr package.

> SATS\_Interest\_NoMissing\_42vs13 <- mutate(SATS\_Interest\_NoMissing, "Sum\_42" = (I29+I20), 'Sum\_13' = (I12+I23))



Once the additional columns have been computed, the cor() function can be used to obtain the correlation between the two average columns.

> cor(SATS\_Interest\_NoMissing\_42vs13[,5:6])

Sum\_42 Sum\_13

Sum\_42 1.0000000 0.7353984

Sum\_13 0.7353984 1.0000000

In this example, the split-half correlation is rsplit-half = .74. One problem with the split-half method is that reducing the number of items on a survey (or test) generally reduces the reliability. Note that each of our split-half versions has only half the items that the full test has for measuring “interest.” To correct for this, you should apply the Spearman-Brown correction:



Task:

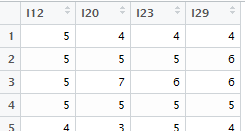
Compute the average inter-item correlation and split-half reliability measure for Effort questions, i.e. questions that start with E. How do these measures of internal consistency compare against those found for Interest? Discuss.

Cronbach’s Alpha

One criticism of the split-half method is that the reliability estimate depends on which “split-halves” are used. To overcome this problem, one could compute the Spearman-Brown corrected split-half reliability coefficients for all possible split-halves and then find their mean.

|  |  |  |
| --- | --- | --- |
| All Possible Combinations | | |
| > sample( c("I12","I20","I23","I29"),replace=F)  [1] "I29" "I20" "I12" "I23"  ( I29 , I20 ) vs ( I12 , I23 ) | > sample( c("I12","I20","I23","I29"),replace=F)  [1] "I20" "I12" "I29" "I23  ( I20 , I12 ) vs ( I29 , I23 ) | > sample( c("I12","I20","I23","I29"),replace=F)  [1] "I12" "I29" "I23" "I20  ( I12 , I29 ) vs ( I23 , I20 ) |
| Split-half correlation  Sum\_42 Sum\_13  Sum\_42 1.0000000 0.7353984  Sum\_13 0.7353984 1.0000000 | Split-half correlation  Sum\_21 Sum\_43  Sum\_21 1.0000000 0.7378283  Sum\_43 0.7378283 1.0000000 | Split-half correlation  Sum\_14 Sum\_23  Sum\_14 1.0000000 0.7961845  Sum\_23 0.7961845 1.0000000 |
| Spearman-Brown | Spearman-Brown | Spearman-Brown |
|  | | |

Cronbach’s α can be computed in R using the alpha() function found in the psych package. The data.frame being passed into the alpha() function should only contain the scores for each of the various items.



The R output for Cronbach’s α is provided here.

> library(psych)

> alpha(SATS\_Interest\_NoMissing)

Reliability analysis

Call: alpha(x = SATS\_Interest\_NoMissing)

raw\_alpha std.alpha G6(smc) average\_r S/N ase mean sd

0.86 0.86 0.84 0.61 6.3 0.011 4.8 1.1

lower alpha upper 95% confidence boundaries

0.84 0.86 0.88

Reliability if an item is dropped:

raw\_alpha std.alpha G6(smc) average\_r S/N alpha se

I12 0.88 0.88 0.83 0.71 7.4 0.0099

I20 0.80 0.80 0.75 0.57 4.0 0.0175

I23 0.80 0.81 0.76 0.58 4.2 0.0168

I29 0.80 0.81 0.75 0.58 4.2 0.0166

Item statistics

n raw.r std.r r.cor r.drop mean sd

I12 429 0.76 0.75 0.61 0.57 4.6 1.4

I20 429 0.88 0.88 0.83 0.77 4.7 1.3

I23 429 0.86 0.87 0.82 0.75 4.9 1.3

I29 429 0.87 0.87 0.83 0.75 5.0 1.3

Non missing response frequency for each item

1 2 3 4 5 6 7 miss

I12 0.03 0.06 0.08 0.29 0.28 0.17 0.10 0

I20 0.03 0.03 0.08 0.25 0.36 0.16 0.08 0

I23 0.02 0.03 0.07 0.17 0.40 0.21 0.10 0

I29 0.02 0.03 0.07 0.19 0.33 0.22 0.15 0

Question:

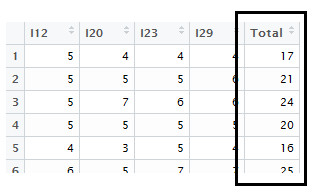
1. What can be said about the internal consistency reliability of the Interest item on the SATS survey? Discuss.

Though the computations above provide insight into what Cronbach’s alpha is measuring, it is never calculated this way in practice. Instead, the following formulas are typically used:

|  |  |
| --- | --- |
| Item Matrix | Breaking down VAR(Items) into   * individual item variances * all pairwise covariances   > ItemVar <- var(SATS\_Interest\_NoMissing[,1:4])  >#Individual variances > ItemVar  I12 I20 I23 I29  I12 1.9468989 0.9779099 0.9148258 0.8889125  I20 0.9779099 1.6976124 1.1178137 1.2905856  I23 0.9148258 1.1178137 1.5710847 1.2147409  I29 0.8889125 1.2905856 1.2147409 1.8188680  ># all pairwise covariances  > ItemVar  I12 I20 I23 I29  I12 1.9468989 0.9779099 0.9148258 0.8889125  I20 0.9779099 1.6976124 1.1178137 1.2905856  I23 0.9148258 1.1178137 1.5710847 1.2147409  I29 0.8889125 1.2905856 1.2147409 1.8188680 |

Next, computing a total score across these four items.

> SATS\_Interest\_NoMissing\_withTotal <- mutate(SATS\_Interest\_NoMissing,"Total" = (I12+I20+I23+I29))



Variance/covariance matrix for Items matrix

The variance in the total score is comprised of two components.

The formula for Cronbach’s α is based on the ratio of these two components.

Wiki Webpage for Cronbach’s Alpha: <https://en.wikipedia.org/wiki/Cronbach%27s_alpha>

Formula #1: 

where *K* is the number of components (items), is the variance of the observed total scores, and is the variance of the scores on the *i*th item.  
  
Getting this computed in R “by hand”…

> (4/3)\*( 1- ( sum(diag(ItemVar)) / var(SATS\_Interest\_NoMissing\_withTotal[,5]) ) )

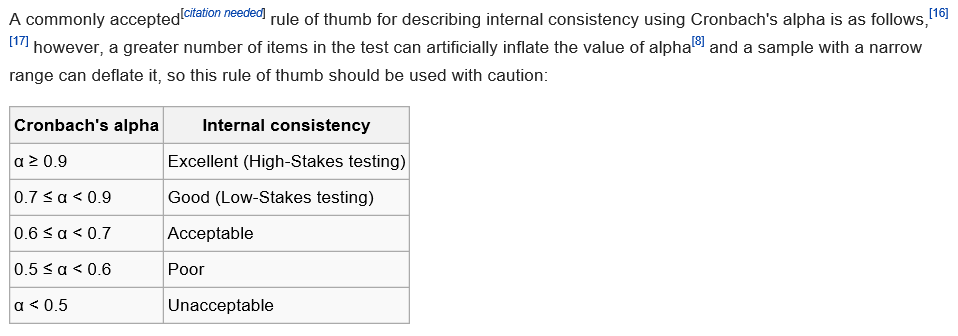
[1] 0.8606834

An alternative measure is **Standardized Cronbach’s alpha** and should be used if the items are measured on different scales.

Formula #2: 

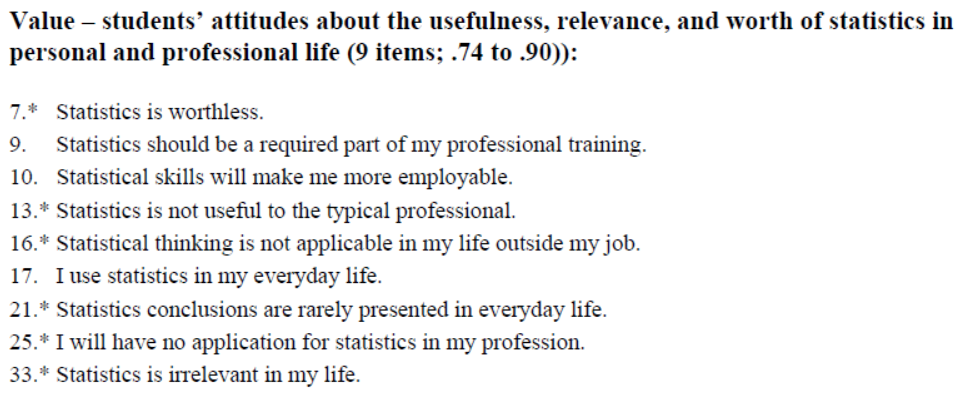
where *K* is the number of components (items), is the average variance of each item, and is the mean of the correlation coefficients.

Comments:

Interpreting Cronbach’s Alpha  
  
To interpret Cronbach’s alpha, consider the following discussion from wikipedia.org.  
  


* There are some problems with this somewhat arbitrary rule of thumb. For example, Cronbach’s alpha tends to increase with the number of items on the scale; so, a high alpha doesn’t necessarily that the measure is “reliable.”
* Cronbach’s alpha is *not* a test for unidimensionality.

Some Words of Caution

Note that Cronbach’s alpha is affected by reverse worded items. For example, consider another component measured in the SATS discussed in above. This survey also measures students’ perceived “value” of statistics using the following questions.  
  


If you don’t value statistics, you might answer Question 7 with “strongly agree” which is coded as a 7. On the other hand, you might answer Question 9 with “strongly disagree” which is coded as a 1. If you have reverse worded items, then you also have to reverse the way in which they are scored before you conduct a reliability analysis. The creators of the SATS include the following instructions on scoring their survey.

