Compute Kappa’s coefficient for inter-rater reliability using StatXact, online calculator, and JMP

Chong Ho Yu, Ph.D. (2013)
Azusa Pacific University
chonghoyu@gmail.com
http://www.creative-wisdom.com/pub/pub.html

This brief write-up demonstrates how Kappa’s coefficient can be computed by StatXact, online calculator, and JMP. There is no single best approach and you are encouraged to try out all of them before deciding which one should be adopted.

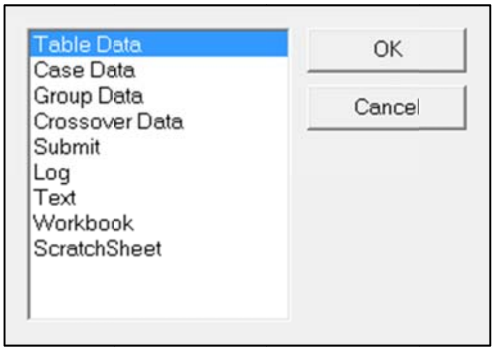
StatXact

This example is based on the data set from Fleiss (1981). In this example, 100 clients were diagnosed by two health care professionals. The subjects were classified into three categories. Obviously, these two experts did not totally agree with each other. For example, Row 1/Column 2 indicates that one client was diagnosed as “neurological” by Rater 1, but the same person was considered “psychological” by Rater 2 (see the yellow highlight).

			Rater 1		
			Column 1	Column 2	Column 3
			Psychological	Neurological	Organic
Rater 2	Row 1	Psychological	75	1	4
	Row 2	Neurological	5	4	1
	Row 3	Organic	0	0	10

Many statistical software applications are capable of computing the Kappa’s coefficient to indicate inter-rater reliability for categorical data. One of the easiest ways is using StatXact. The procedure is illustrated as follows:

From **File** choose **New**. Next, select **Table Data**. For this example, you need a 3X3 table.



Enter the table as shown in the following.

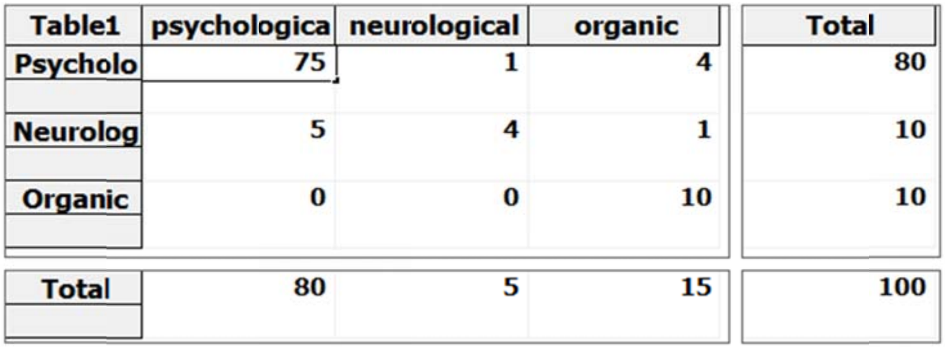
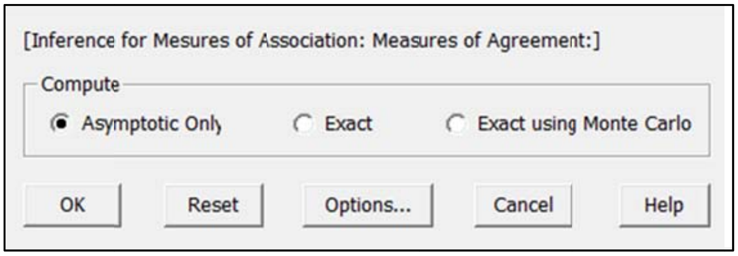


Table1	psychologica	neurological	organic	Total
Psycholo	75	1	4	80
Neurolog	5	4	1	10
Organic	0	0	10	10
Total	80	5	15	100

From **Nonparametrics** choose **Measurement of Agreement** and then **Cohen's Kappa**. You have several options, including the exact test. The exact test is a form of resampling, in which all data are shuffled across different cells in order to simulate chances. It is very resource-intensive. If you are not sure whether your computer has sufficient RAM for this type of processing, it is better to choose "Asymptotic only."

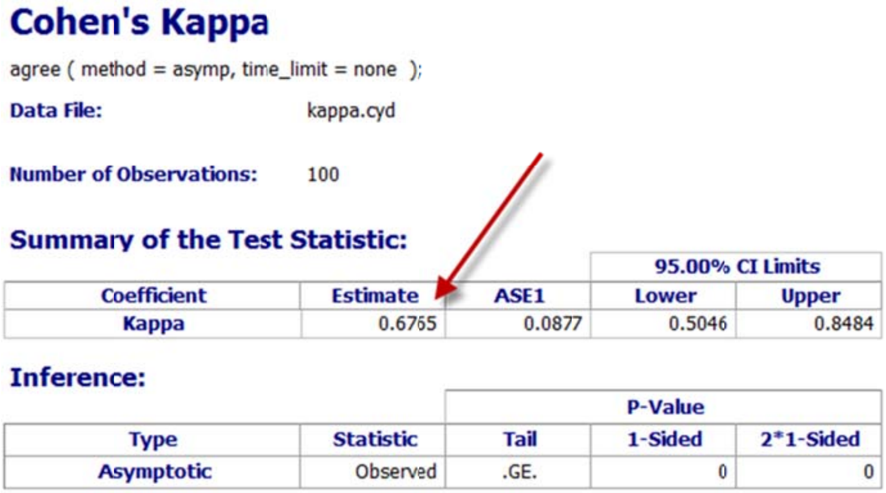


[Inference for Mesures of Association: Measures of Agreement:]

Compute

● Asymptotic Only ○ Exact ○ Exact using Monte Carlo

The results are shown in the following. The Kappa's coefficient is 0.6765. According to Fleiss (1981), kappas over .75 is considered excellent, .40 to .75 is from fair to good, and below .40 is poor. But this is just one of many opinions and currently there is no commonly agreed standard.



Cohen's Kappa

agree (method = asymp, time_limit = none);

Data File: kappa.cyd

Number of Observations: 100

Summary of the Test Statistic:

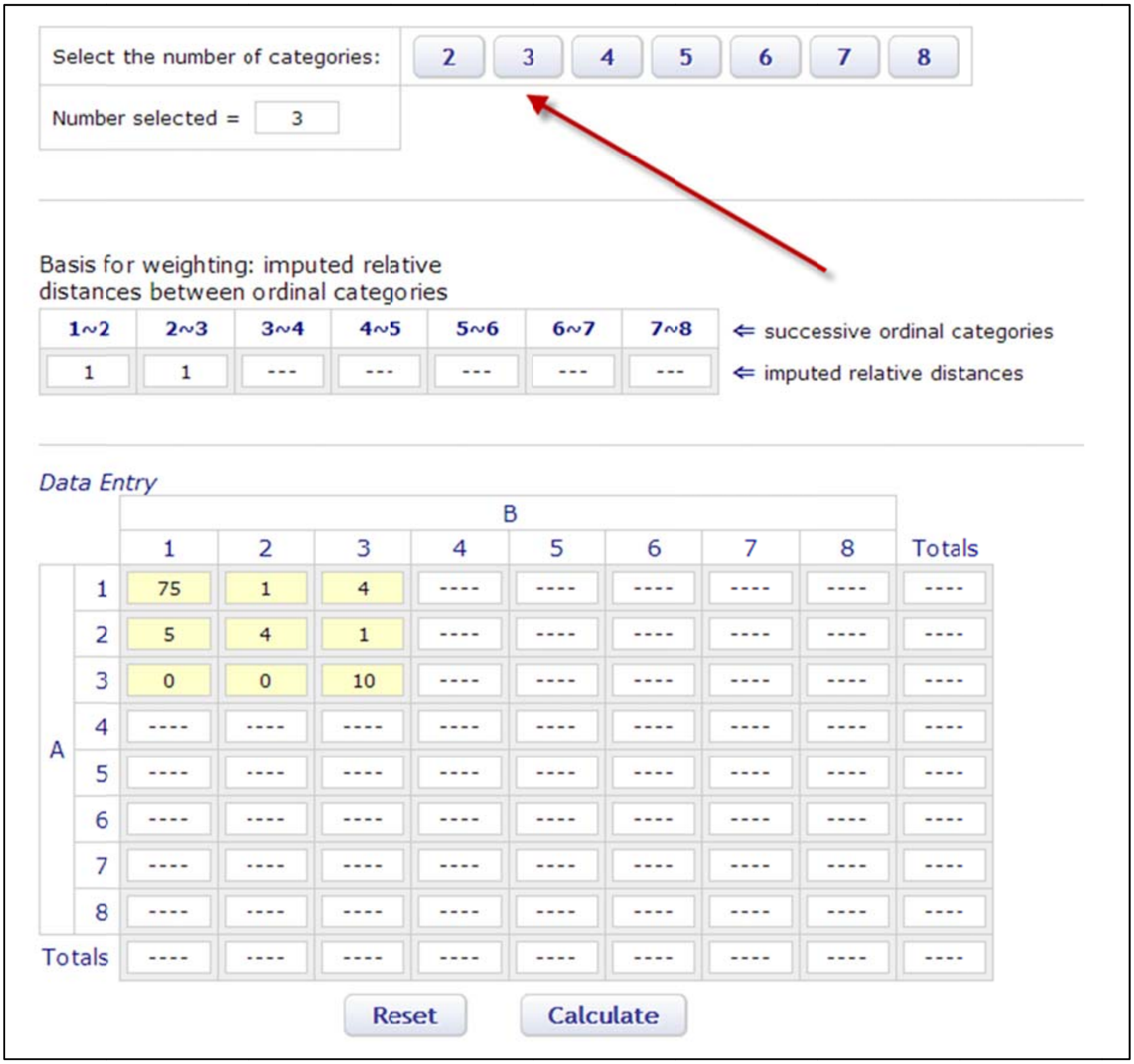
			95.00% CI Limits	
Coefficient	Estimate	ASE1	Lower	Upper
Kappa	0.6765	0.0877	0.5046	0.8484

Inference:

			P-Value	
Type	Statistic	Tail	1-Sided	2*1-Sided
Asymptotic	Observed	.GE.	0	0

Online Kappa calculator

If you have no access to StatXact, another way is to use an online Kappa calculator, such as http://vassarstats.net/kappa.html We can reuse the same data set. First, choose the number of categories by clicking on a button. Next, enter the data into the table (see below). At the end press the Calculate button.



Number selected = 3

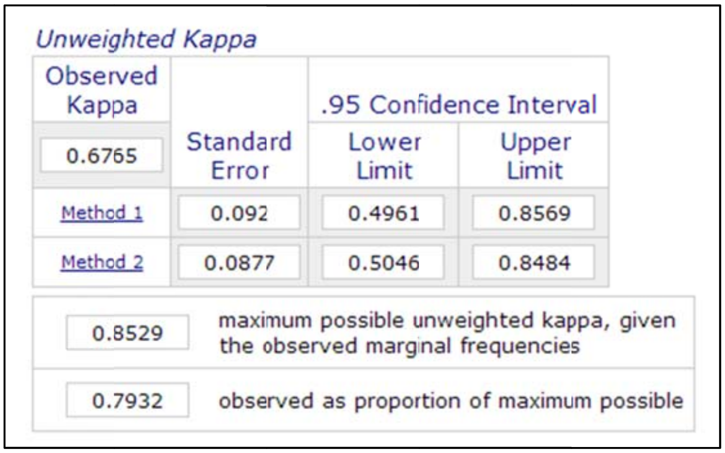
Basis for weighting: imputed relative distances between ordinal categories

1~2	2~3	3~4	4~5	5~6	6~7	7~8	⇐ successive ordinal categories
1	1	---	---	---	---	---	⇐ imputed relative distances

Data Entry

		B								
		1	2	3	4	5	6	7	8	Totals
A	1	75	1	4	----	----	----	----	----	----
	2	5	4	1	----	----	----	----	----	----
	3	0	0	10	----	----	----	----	----	----
	4	----	----	----	----	----	----	----	----	----
	5	----	----	----	----	----	----	----	----	----
	6	----	----	----	----	----	----	----	----	----
	7	----	----	----	----	----	----	----	----	----
	8	----	----	----	----	----	----	----	----	----
Totals		----	----	----	----	----	----	----	----	----

If you do not assign weights to different categories, the user can simply report the unweighted Kappa.



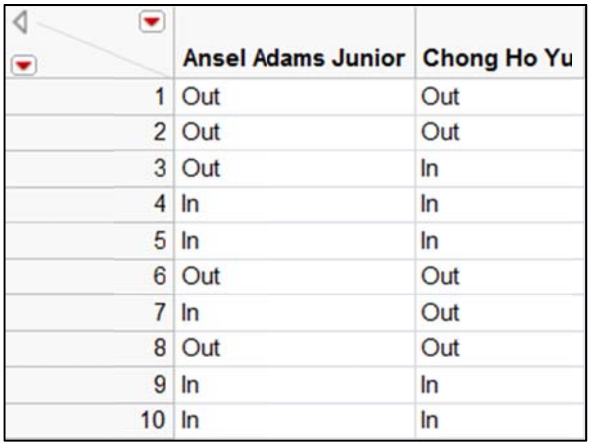
Unweighted Kappa

Observed Kappa		.95 Confidence Interval	
0.6765	Standard Error	Lower Limit	Upper Limit
Method 1	0.092	0.4961	0.8569
Method 2	0.0877	0.5046	0.8484

0.8529	maximum possible unweighted kappa, given the observed marginal frequencies
0.7932	observed as proportion of maximum possible

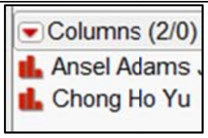
JMP

The preceding two approaches are based upon summary data (the frequency counts have been placed into a row X column table). If you have raw data, it is more convenient to use JMP. Consider this scenario: a photo contest is flooded by many entries. In response to this, the contest organizer hired two photographers (Ansel Adams Junior and Chong Ho Yu) to conduct the first round of screening. In the bale below the entries marked as "In" are selected as the finalists whereas those marked as "Out" are disqualified.



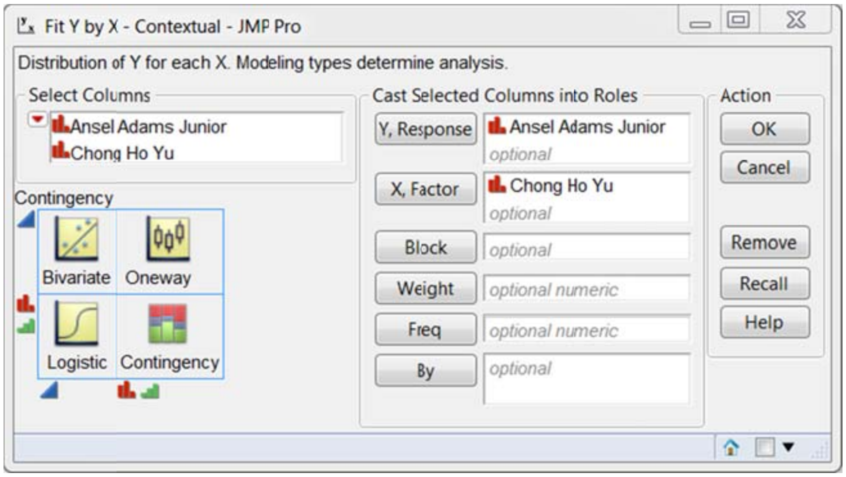
	Ansel Adams Junior	Chong Ho Yu
1	Out	Out
2	Out	Out
3	Out	In
4	In	In
5	In	In
6	Out	Out
7	In	Out
8	Out	Out
9	In	In
10	In	In

If the data are entered as numbers (e.g. 1=In, 0=out), please make sure that they are formatted as "nominal" or "ordinal", not "continuous." You can change the symbol of the data type from a blue triangle to a stack of green bars (ordinal) or a set of red bars (nominal). Either nominal or ordinal can work.



Columns (2/0)
Ansel Adams
Chong Ho Yu

From the pull down menu **Analyze**, select **Fit Y by X**. It does not matter which variable is assigned to be Y and which one is assigned to be X. It is a bi-directional relationship and thus the variables are not classified into the categories “dependent” and “independent”.



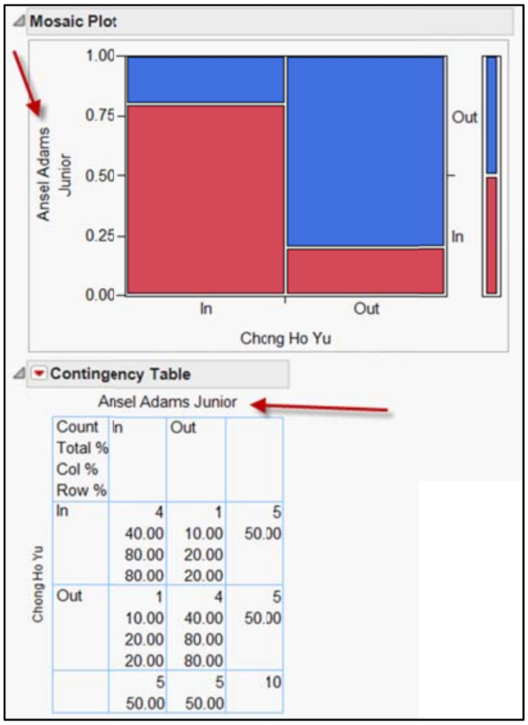
Fit Y by X - Contextual - JMP Pro

Distribution of Y for each X. Modeling types determine analysis.

Y, Response: Ansel Adams Junior

X, Factor: Chong Ho Yu

The first output you can view is the Mosaic plot, which is a graphical version of the crosstab tab. The frequency in each cell is depicted by the size of the area. However, in JMP the orientations of the Mosaic plot and the crosstab table are not consistent. In the plot Ansel’s rating is placed on the Y-axis but in the table it is on the column. You have to mentally rotate the graph or the table to match them. The left red area in the plot depicts the percentage of entries marked as “In” by both Ansel and Chong. This is equivalent to the upper left cell in the table. By looking at the graph alone, you can guess these two judges tend to disagree with each other.



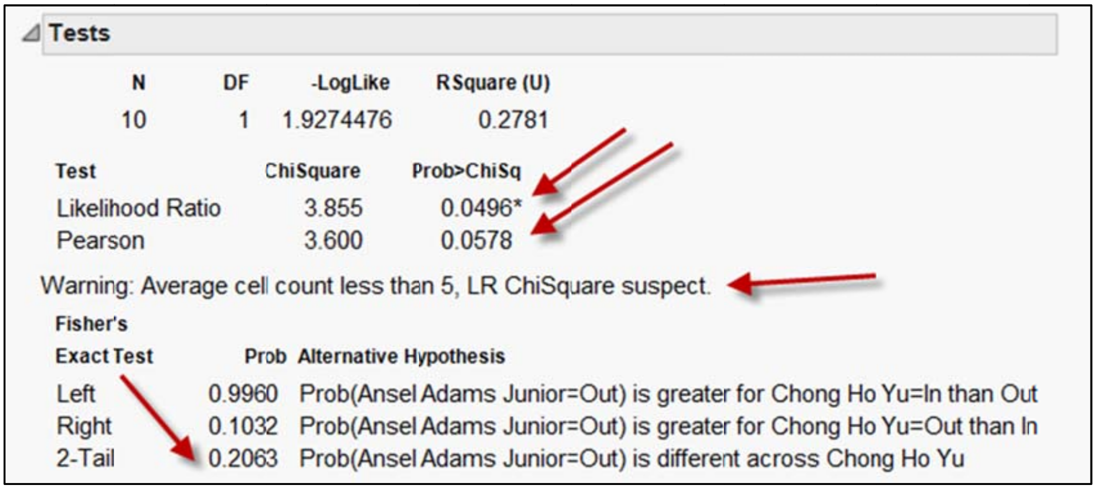
Contingency Table

Ansel Adams Junior

Chong Ho Yu	Count Total % Col % Row %	In	Out	
	In	4 40.00 80.00 80.00	1 10.00 20.00 20.00	5 50.00
	Out	1 10.00 20.00 20.00	4 40.00 80.00 80.00	5 50.00
		5 50.00	5 50.00	10

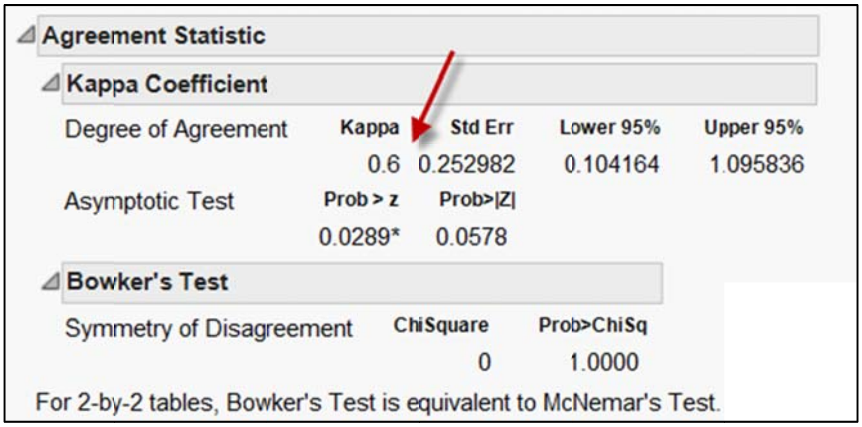
By default JMP returns the Pearson's Chi-square statistics, which is a test of the independence of row and column data. The null hypothesis is: there is no significant association between Ansel's and Chong's ratings. The *p* value of the Pearson's Chi-square is .0578, which is very close to the alpha cut-off. If we adopt the result of the Likelihood ratio (p = .0496), the hypothesis of null or independence is certainly rejected. Does it imply that Ansel's and Chong's ratings are closely related? However, please read the warning: "Average cell count less than 5, likelihood ratio Chi-square suspect." In this situation, any Chi-square-based statistics are suspicious.

Alternatively, we can look at the Fisher's exact test. As mentioned before, the exact test is a form of resampling that takes all possible cell combinations into account. Unlike the Pearson's Chi-square test, the Fisher's exact test yields this two-tailed p value: 0.2063. In other words, the assumption of no association is not rejected, meaning that Ansel and Chong do not go with one accord.



```
Tests

        N      DF     -LogLike     RSquare (U)
       10       1    1.9274476        0.2781

  Test                ChiSquare    Prob>ChiSq
  Likelihood Ratio        3.855       0.0496*
  Pearson                 3.600       0.0578

Warning: Average cell count less than 5, LR ChiSquare suspect.
  Fisher's
  Exact Test          Prob  Alternative Hypothesis
  Left              0.9960  Prob(Ansel Adams Junior=Out) is greater for Chong Ho Yu=In than Out
  Right             0.1032  Prob(Ansel Adams Junior=Out) is greater for Chong Ho Yu=Out than In
  2-Tail            0.2063  Prob(Ansel Adams Junior=Out) is different across Chong Ho Yu
```

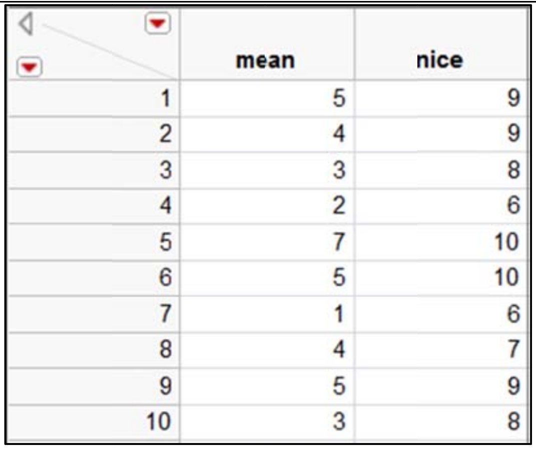
To settle down the dispute, choose **Agreement statistic** from the red triangle. The Kappa coefficient is 0.6. Is it good enough? No! There is a 50% chance that both Ansel and Chong would agree with each other, and thus 0.6 is just a bit better than happening by chance alone.



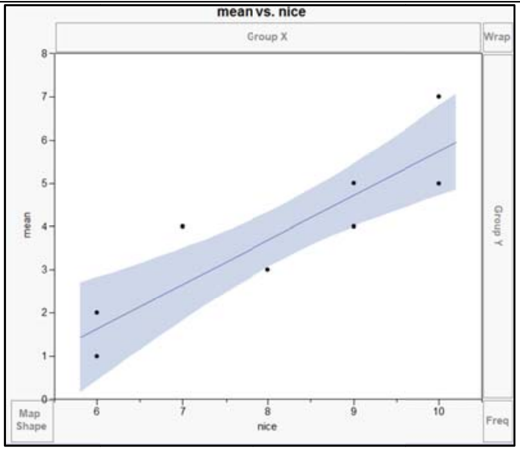
```
Agreement Statistic
  Kappa Coefficient
    Degree of Agreement    Kappa    Std Err    Lower 95%    Upper 95%
                             0.6   0.252982     0.104164     1.095836
    Asymptotic Test      Prob > z    Prob>|Z|
                          0.0289*     0.0578
  Bowker's Test
    Symmetry of Disagreement    ChiSquare    Prob>ChiSq
                                        0        1.0000
 For 2-by-2 tables, Bowker's Test is equivalent to McNemar's Test.
```

Other types of measurement of agreement

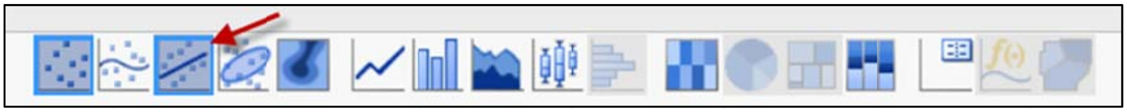
The preceding methods are applicable to nominal data. When the data type are continuous in nature (e.g. scale of 1-10), other types of inter-rater reliability coefficients, such as index of inconsistency (http://www.creative-wisdom.com/teaching/WBI/memory.shtml), intra-class correlation or Pearson's r, are more suitable. However, it is crucial to point out that sometimes the coefficient alone might be misleading. If I tell you that the Pearson's *r* of Rater A's scores and Rater B's scores is .8797, what will be your natural response? You may say, "Wow! High coefficient! The two raters tend to agree with each other. We can trust the panel." Let's look at the data and the scatterplot plot below. These two raters are Mr. Mean and Ms. Nice. Actually, Mr. Mean is a tough grader and his highest score is 7. In contrast, Ms. Nice is very kind to her fellow photographers, and as a result most of them received 8 or above. Obviously, there is a huge discrepancy between the two graders in terms of their perception of the picture quality.



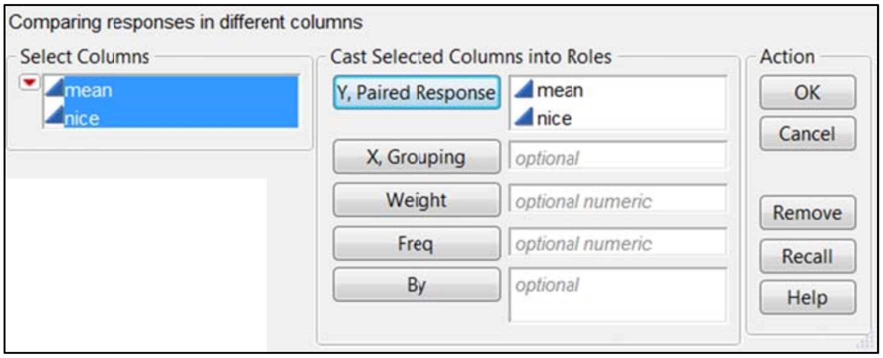
	mean	nice
1	5	9
2	4	9
3	3	8
4	2	6
5	7	10
6	5	10
7	1	6
8	4	7
9	5	9
10	3	8



To create a scatterplot plot in JMP, choose **Graph Builder** from the pull down menu **Graph**. Then drag the scores of the two judges into the Y-axis and X-axis. By default JMP displays a non-linear fit. You can change it to a linear fit by clicking on the **Linear Fit** icon.



To obtain the Pearson's *r*, select **Matched Pairs** from **Analyze**. Put both variables into Y and press OK.



Comparing responses in different columns

Select Columns: mean, nice

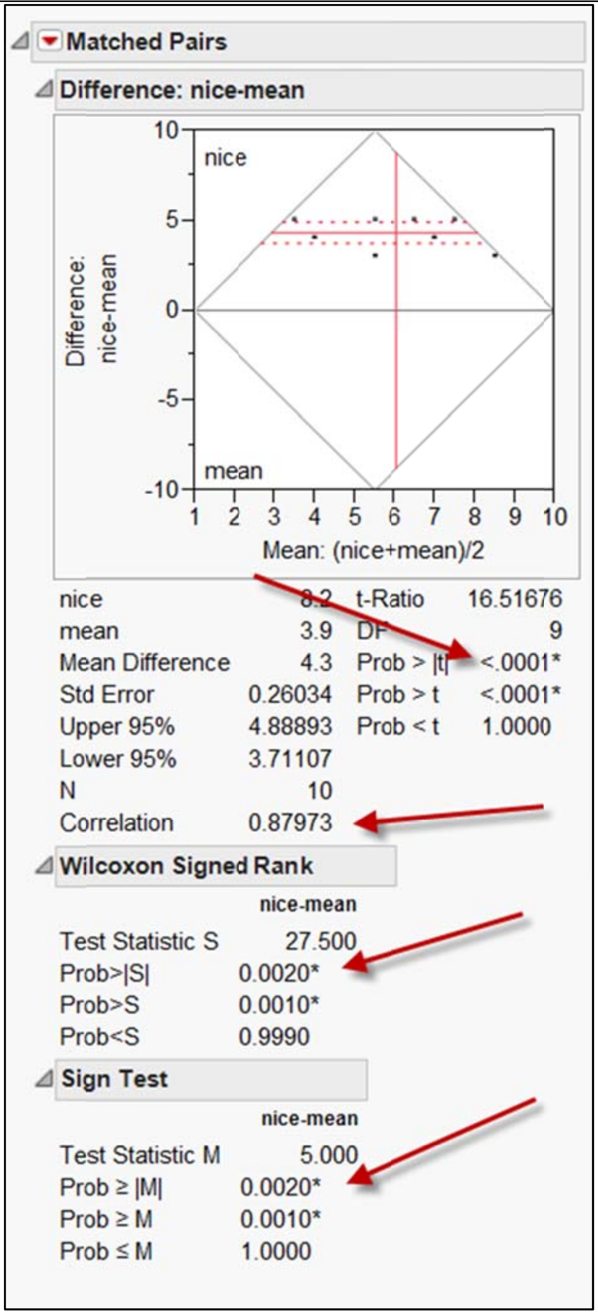
Cast Selected Columns into Roles

- Y, Paired Response: mean, nice
- X, Grouping: optional
- Weight: optional numeric
- Freq: optional numeric
- By: optional

The panel on the right hand side is the output of Matched Pairs. The correction coefficient, as mentioned before, is as high as 0.87973. However, despite the high coefficient, the two raters have vastly different judgments. As their names imply, one is mean and the other is nice.

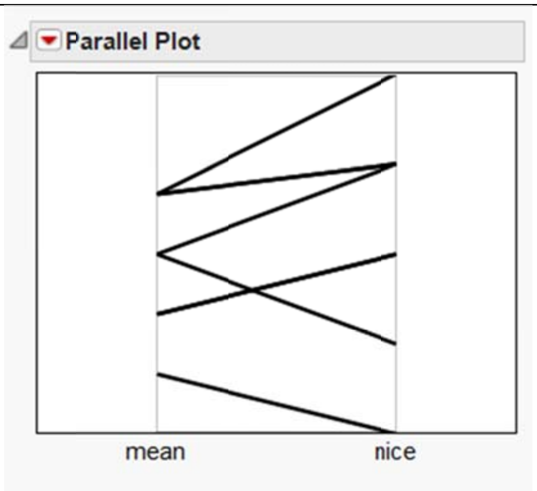
In addition to the correlation coefficient, you need to look at another statistics: two correlated-sample t-test. The t-test indicates whether there is a significant difference between the two mean scores. Not surprisingly, the two-tailed *p* value is .0001, meaning that the null hypothesis is rejected. In other words, on the average the rating of Ms. Nice is much higher than that of Mr. Mean.

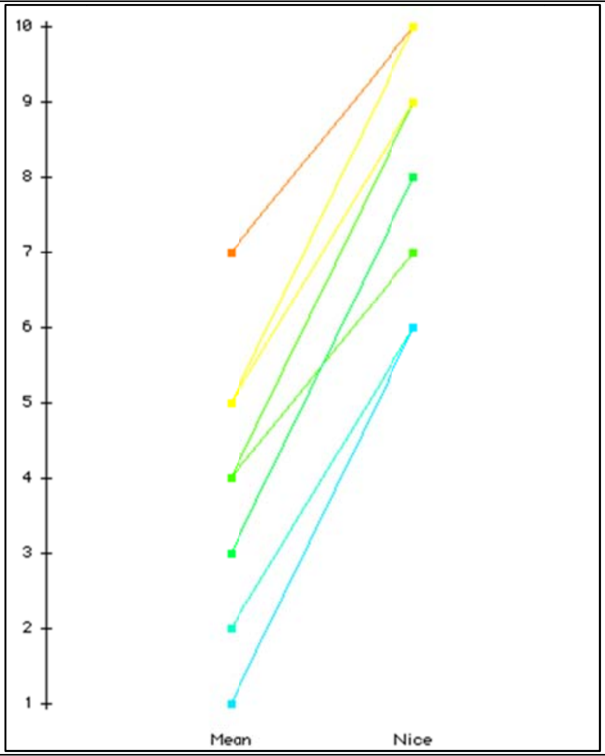
One may argue that these data are not truly continuous (there are no 8.1, 8.2, 8.3, 9.1, 9.2, 9.3, 9.4…etc.). At most we can call them ordinal. If you want to treat the data as rank-ordered, you can request non-parametric tests, such as Wilcoxon Signed Rank test and Sign test, from the **inversed red triangle**. In this example, both non-parametric tests concur with the t-test. Again, it is concluded that the scores of the two judges are significantly different from each other.



```
Matched Pairs
Difference: nice-mean

nice                 8.2    t-Ratio    16.51676
mean                 3.9    DF                9
Mean Difference      4.3    Prob > |t|  <.0001*
Std Error        0.26034    Prob > t    <.0001*
Upper 95%        4.88893    Prob < t     1.0000
Lower 95%        3.71107
N                     10
Correlation      0.87973

Wilcoxon Signed Rank
                     nice-mean
Test Statistic S      27.500
Prob>|S|             0.0020*
Prob>S               0.0010*
Prob<S               0.9990

Sign Test
                     nice-mean
Test Statistic M       5.000
Prob ≥ |M|           0.0020*
Prob ≥ M             0.0010*
Prob ≤ M             1.0000
```

An easy way to show the discrepancy between Mr. Mean and Ms. Nice is using the **Parallel-coordinate Plot** (PCP). However, JMP could not display PCP correctly, as shown in the left panel below. Thus, this author employs DataDesk instead. DataDesk clearly shows an upward trend when the two sets of data points are connected. To create a PCP in DataDesk, choose **Dotplot side by side** from **Plot**, and then select **Lines-Show lines** from **Modify**.





Another possible pitfall of counting on the coefficient alone is that even if the data are truly continuous, the variance may be very low. For example, in the Olympic gymnastics game, all participants are the best of the best, and as a result the dispersion of the scores may be minimal (e.g. 9.7, 9.8, 9.9, 9.7, 9.85, 9.9, 9.67…etc.). The take home message is: always inspect the data by visualization.

References

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed.). New York: John Wiley.

Contact info:
Chong Ho Yu, Ph.D.
chonghoyu@gmail.com
www.creative-wisdom.com