Application of Correlation Analysis: Consistency of Judges

Suppose there are several raters (judges) who will evaluate a number of objects. For example, the objects could be

- contestants in a beauty contest
- figure skaters in the Olympics
- flowers in a garden show
- cookies made from a new recipe

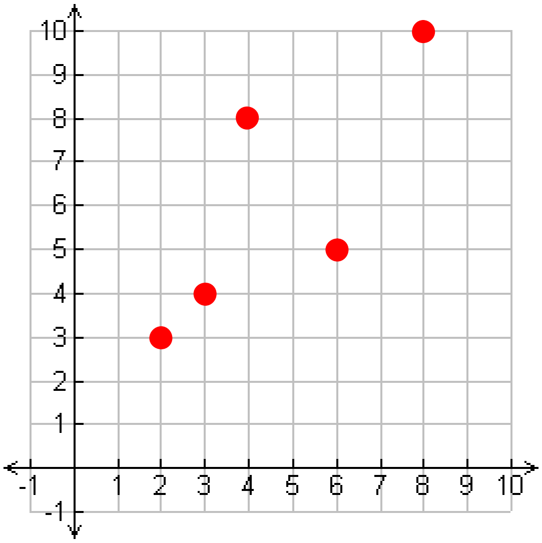
Many of these types of evaluations are subjective, and we are interested in assessing whether the judges' ratings are consistent, or if there are one (or more) judges whose evaluations are somehow different from the others.

To illustrate, suppose there are only two judges, and they each evaluate 5 objects. For each object, each judge assigns a number between 1 and 10, where 1 is least desirable and 10 is most desirable. The judges ratings are given in the following table.

	Object 1	Object 2	Object 3	Object 4	Object 5
Judge 1	2	6	3	4	8
Judge 2	3	5	4	8	10

A graph may help put this data in perspective.

The (x, y) pairs are (judge 1, judge 2) for each object: (2, 3), (6, 5), (3, 4), (4, 8), and (8, 10)



If the two judges have similar ratings for the objects, then these points should be (approximately) on a line.

We can use correlation to measure the strength of the linear relationship.

Correlation in SAS

```
options ls=78 nodate nonumber;

data example;
 input one two;
 title 'Two Judges';
 datalines;
 2 3
 6 5
 3 4
 4 8
 8 10
 ;

proc corr data=example;
run;
```

```
                              Two Judges

                          The CORR Procedure

                     2  Variables:    one       two


                          Simple Statistics

Variable          N        Mean      Std Dev          Sum      Minimum      Maximum

one               5     4.60000      2.40832     23.00000      2.00000      8.00000
two               5     6.00000      2.91548     30.00000      3.00000     10.00000


                 Pearson Correlation Coefficients, N = 5
                        Prob > |r| under H0: Rho=0

                                 one            two

                     one       1.00000        0.78332
                                               0.1171

                     two       0.78332        1.00000
                                0.1171
```

The default in SAS is to calculate Pearson's product-moment correlation coefficient. This is defined as

$$r = \frac{SS_{XY}}{\sqrt{SS_{XX}SS_{YY}}} = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \cdot \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}}$$

Pearson's correlation measures the strength of the linear relationship between the judge's scores. If one judge gives a wider range of scores than the other judge, then the relationship might not be linear. In this case, would be advantageous to use a different method for calculating the correlation. This is a nonparametric alternative to Pearson's method. It is based on ranks, and is called Spearman's correlation coefficient.

Calculating Spearman's correlation coefficient 'by hand'.

	Object 1	Object 2	Object 3	Object 4	Object 5
Judge 1	2 [1]	6 [4]	3 [2]	4 [3]	8 [5]
Judge 2	3 [1]	5 [3]	4 [2]	8 [4]	10 [5]

Convert the actual scores to ranks. For each judge, the smallest value has rank 1, the second smallest value has rank 2, etc. [The ranks are in red. Note that both judges ranked Object 1 the lowest, Object 3 the second lowest, and Object 5 the highest. The judges differed on Objects 2 and 4.]

Then use the formula for calculating r, but use the ranks instead of the actual values.

We can also do this in SAS.

```
title 'Spearman Correlation';
proc corr data=example spearman;
run;
```

```
                           Spearman Correlation

                              The CORR Procedure

                       2  Variables:    one      two


                              Simple Statistics

Variable         N        Mean     Std Dev      Median     Minimum     Maximum

one              5     4.60000     2.40832     4.00000     2.00000     8.00000
two              5     6.00000     2.91548     5.00000     3.00000    10.00000


                  Spearman Correlation Coefficients, N = 5
                         Prob > |r| under H0: Rho=0

                                     one           two

                       one       1.00000       0.90000
                                                0.0374

                       two       0.90000       1.00000
                                  0.0374
```

Spearman's correlation measures the monotone relationship between the judges' scores.
(Monotone means as one value goes up, the other value also goes up. This is not as strict as linear.)

Another example

Five judges evaluate six projects. Assess the consistency of the judges.

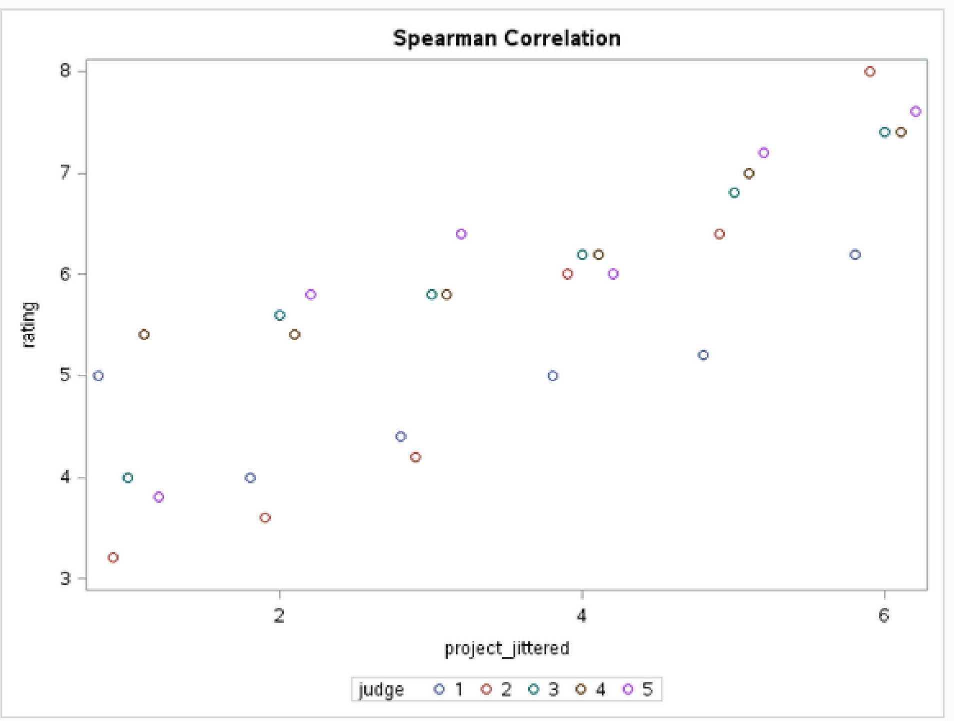
	Object 1	Object 2	Object 3	Object 4	Object 5	Object 6
Judge 1	5.0	4.0	4.4	5.0	5.2	6.2
Judge 2	3.2	3.6	4.2	6.0	6.4	8.0
Judge 3	4.0	5.6	5.8	6.2	6.8	7.4
Judge 4	5.4	5.1	5.8	6.2	7.0	7.4
Judge 5	3.8	5.8	6.4	6.0	7.2	7.6

First read the data into SAS.

```
data panels;
 input judge project rating @@;
 project_jittered = project + (judge -3)/10;
    * to jitter the points, so they don't overlap on the graph;
 datalines;
 1 1 5.0  1 2 4.0  1 3 4.4  1 4 5.0  1 5 5.2  1 6 6.2
 2 1 3.2  2 2 3.6  2 3 4.2  2 4 6.0  2 5 6.4  2 6 8.0
 3 1 4.0  3 2 5.6  3 3 5.8  3 4 6.2  3 5 6.8  3 6 7.4
 4 1 5.4  4 2 5.4  4 3 5.8  4 4 6.2  4 5 7.0  4 6 7.4
 5 1 3.8  5 2 5.8  5 3 6.4  5 4 6.0  5 5 7.2  5 6 7.6
 ;
```

Then generate a plot. Note that there are five judges this time, so the plot is different.

```
proc sgplot data=panels;
 scatter x=project_jittered y=rating /group=judge;
 run;
```



Next we calculate the correlation. Since correlation is only defined for pairs of values, we will need to calculate the correlation between each pair of judges. We calculate both Pearson and Spearman correlations.

```
data pairwise;
input project judge1 judge2 judge3 judge4 judge5;
datalines;
1 5.0 3.2 4.0 5.4 3.8
2 4.0 3.6 5.6 5.4 5.8
3 4.4 4.2 5.8 5.8 6.4
4 5.0 6.0 6.2 6.2 6.0
5 5.2 6.4 6.8 7.0 7.2
6 6.2 8.0 7.4 7.4 7.6
;

proc corr data=pairwise pearson spearman;
 var judge1 -- judge5;
 run;
```

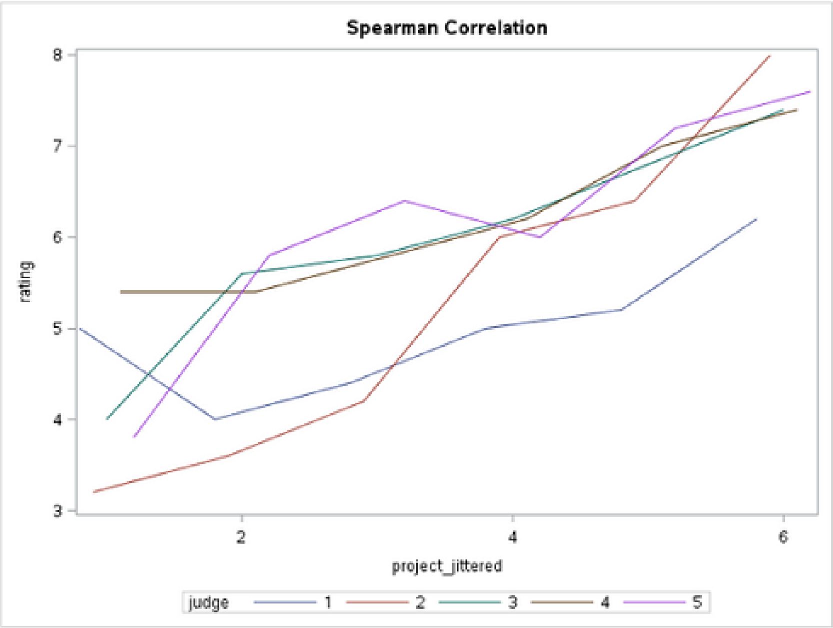
```
                                 New Example

                              The CORR Procedure

         5  Variables:    judge1   judge2   judge3   judge4   judge5


                              Simple Statistics

Variable         N        Mean     Std Dev      Median     Minimum     Maximum

judge1           6     4.96667     0.75277     5.00000     4.00000     6.20000
judge2           6     5.23333     1.86940     5.10000     3.20000     8.00000
judge3           6     5.96667     1.16905     6.00000     4.00000     7.40000
judge4           6     6.20000     0.83905     6.00000     5.40000     7.40000
judge5           6     6.13333     1.33666     6.20000     3.80000     7.60000


                   Pearson Correlation Coefficients, N = 6
                          Prob > |r| under H0: Rho=0

               judge1        judge2        judge3        judge4        judge5

 judge1       1.00000       0.82526       0.53483       0.83596       0.42669
                             0.0431        0.2742        0.0382        0.3988

 judge2       0.82526       1.00000       0.90479       0.96907       0.81747
               0.0431                      0.0132        0.0014        0.0469

 judge3       0.53483       0.90479       1.00000       0.88084       0.97614
               0.2742        0.0132                      0.0205        0.0008

 judge4       0.83596       0.96907       0.88084       1.00000       0.82745
               0.0382        0.0014        0.0205                      0.0421

 judge5       0.42669       0.81747       0.97614       0.82745       1.00000
               0.3988        0.0469        0.0008        0.0421


                  Spearman Correlation Coefficients, N = 6
                          Prob > |r| under H0: Rho=0

               judge1        judge2        judge3        judge4        judge5

 judge1       1.00000       0.75370       0.75370       0.83824       0.66674
                             0.0835        0.0835        0.0371        0.1481

 judge2       0.75370       1.00000       1.00000       0.98561       0.94286
               0.0835                      <.0001        0.0003        0.0048

 judge3       0.75370       1.00000       1.00000       0.98561       0.94286
               0.0835        <.0001                      0.0003        0.0048

 judge4       0.83824       0.98561       0.98561       1.00000       0.92763
               0.0371        0.0003        0.0003                      0.0077

 judge5       0.66674       0.94286       0.94286       0.92763       1.00000
               0.1481        0.0048        0.0048        0.0077
```

Are the Five Judges Consistent?

Two judges are consistent if the correlation between their ratings is close to 1. This is true for all judges except judge 1. If we use Pearson's correlation (the top matrix on page 6), notice that all correlations are above 0.8 except for judges 1 and 3 and for judges 1 and 5. So judge 1 seems to be different than judges 3 and 5, but the other four judges seem to be consistent. If we use Spearman's correlation (the bottom matrix on page 6), we arrive a similar conclusion. All of the Spearman correlations are above 0.8 except for judge 1 with judge 2, 3 or 5. Again, it seems that judge 1 assigns different ratings than the other judges.

We can also arrive at the same conclusion by looking at the p-values provided in the correlation matrix. These p-values are for testing H_0: $\rho = 0$ vs. H_a: $\rho \neq 0$. Consistent judges will have ρ close to 1, so we expect to reject the null hypothesis if the judges are providing consistent ratings. Using Pearson's correlation, this is not the case for judge 1 and judge 3 ($p = 0.2742$) or for judge 1 and judge 5 ($p = 0.3988$). In both of these pairings, we fail to reject the null hypothesis that the correlation is zero.

For this particular data set, it seems that judge 1 is not consistent with the other judges.



This graph is the same as the one on page 45, but I have added lines connecting the points for each judge. This makes it easier to see the "profile" for each judge, and can sometimes help expose inconsistencies. In this graph, we see that judge 1 (whom we have already identified as being "different") has lower ratings for most projects except for project 1. All of the other judges have increasing ratings across the projects (except for a slight dip for project 4 and judge 5). Judge 1 is different from the other judges primarily because of the rating for project 1, which appears to be higher than we would expect if all the judges were consistent.