24. Inter-rater Reliability

Researchers will often hire staff to gather their data. This can take many forms, including structured interviews, administering a scale or pre-determined set of questions, or observing and recording specific behaviors. Students doing a study of who comes to a complete stop needed to be sure that everyone in their group could consistently apply the definition of complete stop. These question arise in experimental design, and they all have to do with inter-rater Reliability. In this section, we'll discus how to test for this when the research instrument is a scale.

24.1. Example.

A researcher is interested in how a scale that screens for substance abuse that is used in different environments: in an Emergency Room, a physician's office, and a DHS office. The researcher gathered data of ten subjects from each setting, with the data shown at right.
The researcher is interested in whether the scores depend on location and whether the scores differentiate between subjects.

	ER	Physician Clinic	DHS Clinic
	1	3	1
	2	9	2
	1	7	1
	8	2	2
	1	3	1
	2	2	1
	7	3	2
	5	1	2
	3	3	1
	3	2	1

Solution.

Two Way Analysis of Variance with no replacement

Put your data in the green cells.
Do not edit any other cells.
Do not add or delete columns or rows.
There are hidden columns where most of the calculations are done.

Total Ratings	30
Number of Raters	3
Number of Subjects	10

	SS	dF	MSq	F	p	
SSt	143.8667	29	4.96092			
SSr	29.86667	9	3.318519	0.685539	71.29%	<--tests for whether columns are the same
SSc	26.86667	2	13.43333	2.775057	8.90%	<-- tests for whether the rows are the same
Sse	87.13333	18	4.840741			

Reliability	-0.45871	<--Intraclass correlation

Avererages	3.3	3.5	1.4					
	Groups							
Subjects	A	B	C	D	E	F	G	H
1	1	3	1					
2	2	9	2					
3	1	7	1					
4	8	2	2					
5	1	3	1					
6	2	2	1					
7	7	3	2					
8	5	1	2					
9	3	3	1					

The test for whether the scores are the same or differ by column (location) has a p-value of 71%, so we must conclude the tests are different. This isn't surprising since the average at DHS is so much lower than the other two locations.
The test for whether the row averages are the same has a p-value of 8.9%, so we can conclude that they are different. This means that the scale does differentiate between subjects.
The intra-class correlation describes how well values in the same group correlate with each other. A rough interpretation is that

less that 0.4	poor
between 0.4 and 0.50	fair
between 0.6 and 0.74	good
0.75 and higher	excellent

The spreadsheet calculates the ICC for the columns only. This means there is fair reliability within the locations.

The spreadsheet is essentially the same as the earlier ANOVA spreadsheet, except that identifies an additional source of variability in the sample: the variability due to the individual rows. It still calculates the sum of squares for the columns in the same way, but now it does an ANOVA on the transpose of the rows:

Subjects	1	2	3	4	5	6	7	8	9	10
ER	1	2	1	8	1	2	7	5	3	3
Physician	3	9	7	2	3	2	3	1	3	2
DHS	1	2	1	2	1	1	2	2	1	1

This changes the calculation of the residuals accordingly. However, this is still fundamentally the one-way ANOVA spreadsheet from earlier, slightly tweaked.

The other added feature is the intra-class correlation, described in the above problem. It's calculated in a manner similar to the correlation coefficient.

24.2. Example.

The Human Resources Department at Mechanics R Us uses a standardized set of questions to assess job satisfaction. The VP for Human Resources wonders if the questions produce consistent results between interviewers and if the questions distinguish between different levels of satisfaction.

She randomly selects ten applicants and five interviewers. Each interviewer asks the ten applicants the questions and assigns a score to their responses. The results are at right.

Subjects	A	B	C	D	E
1	100	83	95	88	
2	90	86	94	80	
3	85	84	100	92	
4	85	90	91	95	
5	73	82	64	88	
6	72	74	79	77	
7	72	72	70	75	
8	84	85	90	77	
9	91	100	100	83	
10	81	83	85	91	

The VP's questions can be formulated as two hypotheses:

H_1 : The row averages are different

H_2 : The Column averages are the same

The hope is that the questionnaire $does$ differentiate between the subjects, and thus that our evidence leads us to accept $\boldsymbol{H_1}$. At the same time, the hope is that the interviewers' use of the instrument is $reliable$, i.e., that we do **not** have evidence to reject $\boldsymbol{H_2}$. The second hypothesis is checking for inter-rater reliability.

Solution.

Two Way Analysis of Variance with no replacement

Put your data in the green cells.
Do not edit any other cells.
Do not add or delete columns or rows.
There are hidden columns where most of the calculations are done.

Total Ratings	40
Number of Raters	4
Number of Subjects	10

	SS	dF	MSq	F	p	
SSt	3193.1	39	81.87436			
SSr	1986.1	9	220.6778	5.24083	0.04%	<--tests for whether columns are the same
SSc	70.1	3	23.36667	0.55493	64.93%	<-- tests for whether the rows are the same
Sse	1136.9	27	42.10741			

Reliability 0.809191 <--Intraclass correlation

Avererages	83.3	83.9	86.8	84.6				
	Groups							
Subjects	A	B	C	D	E	F	G	H
1	100	83	95	88				
2	90	86	94	80				
3	85	84	100	92				
4	85	90	91	95				
5	73	82	64	88				
6	72	74	79	77				
7	72	72	70	75				
8	84	85	90	77				
9	91	100	100	83				
10	81	83	85	91				
11								

In this case, we accept the hypothesis that the columns are the same, i.e., that there is no interviewer bias since the p-value is 0.04%.
We reject the hypothesis that the rows are the same, i.e., that the scale does differentiate between the subjects.
The ICC of 0.809 says that the reliability for the interviewers is excellent.

There are two factors or sources of variability that contribute to answering the VP's questions. The first factor is the variability due to the questions themselves. The second factor is the variability due to the subjects. The spreadsheet calculations are similar to those for one-factor ANOVA, except that all the calculations now take into account all the sources of interaction and are thus a bit more complex. In any case, for each factor, we will calculate the Mean-Square Error between the components of the factor and within the components of the factor.