Statistics Cheat Sheet for Inter-Rater Reliability

Tests

Nominal Data:
- Cohen's kappa

Ordinal Data:
- Kendall's tau (recommended)
- Spearman's rank-order correlation coefficient

Interval or Ratio Data:
- Intraclass correlation coefficient (recommended)
- Pearson product-moment correlation coefficient (**not** recommended because it will consistently give an **overestimate** of the degree of agreement between raters)

Cohen's kappa

k= (Oa-Ea)/(N-Ea)

where: K = kappa statistic
Oa = observed count of agreement
Ea = expected count of agreement
N = total number of responses

Generally want values of k greater than 0.70

Example Two doctors classify 100 patients as having either schizophrenia, bipolar disorder, or some other mental illness. The data are **nominal** data. You want to know how well the two doctors agree in their assessment of the patients.

	Doctor 1			**Row Sums**
Doctor 2	**Schizophrenia**	**Bipolar Disorder**	**Other**	
Schizophrenia	31	4	2	**37**
Bipolar Disorder	6	29	8	**43**
Other	10	7	3	**20**
Column Sums	**47**	**40**	**13**	**100**

Ea for R1C1 ` = (37)(47)/100 = 17.39
Ea for R2C2 = (43)(40)/100 = 17.20
Ea for R3C3 = (20)(13)/100 = 2.60

Ea = 17.39 + 17.20 + 2.60 = 37.19
Oa = 31 + 29 + 3 = 63

k = (63 – 37.19) / (100 – 37.19)
= 25.81 / 62.81 = **0.41** This is an **unacceptably low** value of Cohen's kappa.

Intraclass Correlation Coefficient

Use a ***single factor, within-subjects analysis of variance.***

Example. Three judges scores the difficulty of 6 questions for a math test. They gave each question a score between 1 (very easy) and 10 (very difficulty). You want to know how well the three judges agree in their evaluations of the difficulty of the proposed test items. The data are **interval.**

Question	**JUDGE_1**	**JUDGE_2**	**JUDGE_3**
1	9	7	4
2	10	8	7
3	7	5	3
4	10	8	7
5	7	5	2
6	8	6	6

Results of single-factor, within-subjects analysis of variance

Summary of all Effects; design: (intraclass data.sta)
1-RFACTOR1

	df Effect	MS Effect	df Error	MS Error	F	p-level
1	2	20.22222	10	.488889	41.36364	**.000015**

Since the p value is very low (p=.000015), the probability that the differences in scores assigned by the three judges are by chance is also very low. E.g., the three judges **do differ significantly** in their evaluation of the difficulty of the test questions. Some of your questions are not good ones because these three judges differ greatly in terms of how difficult they think the questions are. As a general rule of thumb, p values of less than .30 mean that the level of agreement between raters is poor.

Pearson Product-Moment Correlation

This test is not recommended for evaluating inter-rater reliability because it consistently **overestimates** agreement. Using the same data as the previous example, here are the results of the Pearson product-moment correlation.

Correlations (intraclass data.sta)
Marked correlations are significant at p < .05

	JUDGE_1	JUDGE_2	JUDGE_3
JUDGE_1	1.00	1.00**	0.85**
JUDGE_2	1.00**	1.00	0.85**
JUDGE_3	0.85**	0.85**	1.00

As you can see, unlike the previous test, this test indicates that the correlation (agreement) between all three judges is high – that the chance that the differences are random is low (p < 0.05). This is the problem with this test. While easy to perform, it consistently overestimates agreement.

Spearman Rank-Order Correlation

In this case we have **ordinal** data. Three judges **ranked** how difficult the six test questions are, in their opinion.

QUESTION	JUDGE_1	JUDGE_2	JUDGE_3
1	4	4	3
2	6	5	5
3	1	2	2
4	5	6	6
5	2	1	1
6	3	3	4

Results of Spearman's Rank-Order Correlation Test

	Valid N	Spearman R	t (N-2)	p-level
JUDGE_1 & JUDGE_2	6	.885714	3.815836	.018845
JUDGE_1 & JUDGE_3	6	.828571	2.959800	.041563
JUDGE_2 & JUDGE_3	6	.942857	5.659453	.004805

These results indicate that there is a **high** level of agreement between the three judges for two reasons. First, examine the Spearman R statistic. It is well above 0.70 in every case. Second, look at the p values. They are low in all cases indicating that the probability that the high values for Spearman R occurred by chance is very low.

Kendall's Tau

Kendall's Tau is a **more conservative** test of association. It is probably a better test to use than Spearman's Rank-Order Correlation because it is less likely to **overestimate** the probability that agreement exists.

	Valid N	Kendall Tau	Z	p-level
JUDGE_1 & JUDGE_2	6	.733333	2.066540	.038778
JUDGE_1 & JUDGE_3	6	.600000	1.690806	.090874
JUDGE_2 & JUDGE_3	6	.866667	2.442275	.014595

As you can see, this test gives a lower correlation coefficient in every case, and gives a **higher** p value in every case. So it is **conservative** in terms of both the level of agreement (correlation coefficient) and in terms of the probability of chance agreement (p value).