TRAINER'S Corner

Measurement system analysis with attribute data

By Keith M. Bower, M.S.

A measurement system analysis is a critical component for any quality improvement process. To understand the adequacy of a measurement system, strategies have been developed to assess characteristics such as the proportion of variability contributed by the measurement system to the total variation. The methodology is widely known as Gage Repeatability and Reproducibility[1] (Gage R&R). Guidelines are available from several sources including the Automotive Industry Action Group[2] (AIAG).

Typical Gage R&R analysis requires that the response variable is quantitative (continuous). However, appraisers must frequently use subjective assessments that are qualitative (attribute). For example:

- Quality of cereal bar flavor on a scale of 1 to 10
- Call center operator answered inquiry correctly / incorrectly
- Employee performance rated on a scale of 1 to 3
- Noise exhibited from engine was a squeak, thump, whirr, etc.

MINITAB Release 13.3 provides Attribute Gage R&R to assess a measurement system for responses of this type.

Statistical Analysis

MINITAB includes Attribute Gage R&R functionality to assess a measurement system for qualitative response variables. This paper addresses two statistics: Kappa and Kendall's Coefficient of Concordance. The use of these statistics depends on whether the response data are ordinal or nominal. The following example illustrates the difference in interpretation between these two statistics.

Scenario

Consider a central call center for a major credit card company fielding incoming calls from customers with account inquiries. The Quality Assurance (QA) department conducts regular inspections of their operators by:

- Calling and asking questions similar to those customers may ask. Inspectors may disguise their voice to avoid recognition from the operator.
- Listening to recordings of actual customer calls.

Grades received on characteristics such as friendliness, accuracy, and suitable advice are used to obtain a comprehensive score (1-5). The comprehensive scores are:

1 = perfect response
2 = good response
3 = average response
4 = poor response
5 = very bad response

Because the QA specialists are subjectively rating these operators, and there may be serious consequences for consistently receiving poor scores (4 or 5), you need to ensure that the level of agreement among these QA specialists is high.

Method

In an isolated room, three QA specialists: Judith, Malcolm, and Susan, listen to ten recorded conversations between actual customers and operators. Each specialist grades the operators' responses based on characteristics such as friendliness, accuracy, and suitable advice. These grades are used to obtain a comprehensive score, which is used in the analysis. The results are shown below:

Recording	Judith	Malcolm	Susan
1	1	2	2
2	2	2	2
3	3	3	4
4	5	5	5
5	2	3	3
6	4	4	5
7	1	2	2
8	1	2	2
9	3	2	3
10	4	4	5

Looking across rows, you can see that the level of agreement for each recording is fairly high. Susan always gives the lowest rating of the three appraisers, although she is sometimes in agreement with one or both of the other operators.

To perform this analysis, see the online tutorials page at **http://www.minitab.com/resources/tutorial** for more information on the Attribute Gage R&R functionality in MINITAB.

If you have a release of MINITAB earlier than Release 13.3, you will need to download the most recent maintenance update to get Attribute Gage R&R functionality at: **http://www.minitab.com/support/maintenance/index.htm**

Results

To test the level of absolute agreement among appraisers, researchers often use the Kappa statistic. Kappa ranges between –1 and +1. A Kappa of 1 indicates perfect agreement. Negative values occur when agreement is weaker than expected by chance, which rarely happens. The null hypothesis for this test is that Kappa is equal to zero. That is, the level of agreement among the QA scores per call could be obtained by chance alone.

The output below shows the overall Kappa is 0.2982, with a corresponding p-value of 0.001. You would reject the null hypothesis at the $\alpha = 0.05$ significance level and conclude that the level of agreement is higher than expected by chance.

Between Appraisers

Assessment Agreement

# Inspected	# Matched	Percent (%)	95.0% CI
10	2	20.0	(2.5, 55.6)

Matched: All appraisers' assessments agree with each other.

Kappa Statistics

Response	Kappa	SE Kappa	Z	P(vs > 0)
1	-0.1111	0.1826	-0.6086	0.729
2	0.2823	0.1826	1.5462	0.061
3	0.3750	0.1826	2.0540	0.020
4	0.2800	0.1826	1.5336	0.063
5	0.5200	0.1826	2.8482	0.002
Overall	0.2982	0.0968	3.0799	0.001

Because it is relatively easy to reject the null hypothesis for this test, practitioners typically require Kappa values of 0.7 or above to conclude that the measurement system is adequate. Kappa values above 0.9 are generally regarded as representing very good agreement. For more information and references, see Futrell[3].

In this example, there is low absolute agreement (Kappa = 0.2982) among the QA specialists if you only look at the actual score given. However, because the scores represent a rating (ordinal data), the Kappa statistic does not adequately represent the level of agreement. The Kappa statistic does not take into account the magnitude of the disagreement. In this example, the QA specialists tended to agree with each other, even though they did not give exactly the same score. The level of disagreement is only of the order of 1 ranking at most. The Kappa statistic cannot discern a difference between 1 and 5 as distinct from 1 and 2. To assess the level of agreement when you have a ranking system, use Kendall's Coefficient of Concordance (KCC). KCC, unlike Kappa, accounts for the magnitude of the difference among scores. When KCC ranges from 0 to 1; a higher value indicates stronger agreement. For more information on Kappa and KCC refer to Fleiss[4], and Siegel and Castellan[5], respectively.

As shown below, KCC is 0.9409, indicating a high level of agreement among the ratings of the QA specialists when the magnitude of differences is accounted for.

Kendall's Coefficient of Concordance

Coef	Chi - Sq	DF	P
0.9409	25.4035	9	0.003

Conclusion

This example shows that the correct use of these two statistics is crucial for a valid interpretation of the results from an Attribute Gage R&R study. When there is a ranking system, practitioners are advised to use both Kappa and Kendall's Coefficient of Concordance to better understand the level of agreement among appraisers. ❖

References

1. *Bower, K.M., Touchton, M.E. (April 2001). "Evaluating The Usefulness of Data Using Gage Repeatability and Reproducibility," Asia Pacific Process Engineer.*
2. *Automotive Industry Action Group (February 1995). Measurement System Analysis, 2nd Edition.*
3. *Futrell, D. (May 1995). "When Quality Is a Matter of Taste, Use Reliability Indexes," Quality Progress, Vol. 28, No. 5.*
4. *Fleiss, J.L. (1981). Statistical Methods for Rates and Proportions, 2nd Edition. Wiley & Sons, NY.*
5. *Siegel, S., Castellan, N. J. Jr. (1988). Nonparametric Statistics for the Behavioral Sciences, 2nd Edition. McGraw-Hill, NY.*

Keith M. Bower, M.S., is a technical training specialist with Minitab Inc.

DO YOU KNOW?... **You can run Attribute Gage R&R in Release 13.31 to assess the consistency of classifications when measurements are subjective judgments or ratings made by people.**