Evaluation of Inter-Rater Reliability

by

Matthew Ladew
Clarkson University

Evaluation of Inter-Rater Reliability

A Thesis Proposal by

Matthew Ladew

Department of Electrical Engineering

Mentor: Edward Sazonov

March 12, 2010
Abstract:
In the interest of monitoring a subject’s daily physical activities, a device has been designed to be worn by a subject and to classify and record several common domestic activities. This device will be tested for reliability by comparing the data collected by the device with a standard for each test. The standards will be the result of raters reviewing video recordings of each test and annotating the videos with specially developed software. The reliability between these raters will then be determined through various statistical methods. Comparing these statistics will help determine which methods are more convenient and effective for calculating inter-rater reliability. Ultimately these statistics will validate the standard generated by the raters, and thus the results of the reliability study as a whole.

Introduction:
An outpatient undergoing physical therapy (e.g. a rehabilitating stroke patient) may have a goal for the amount of physical activity he or she performs daily. Such patients and their therapists may desire a reliable means of approximating their activities throughout the day without the patients having to manually record each activity and then periodically present this record to the therapists. To this end, a device has been designed to be worn inside a patient’s shoe and classify the subject’s activities for some period. The refinement of this device is still in progress, and tests need to be performed to determine its reliability in classifying each activity.

Performing a reliability study requires a standard against which to compare the data to be studied. Often there is a test that stands as a gold standard of reliability. In some studies, no such standard exists, and must be generated by human raters. Human raters, however, can be inconsistent and in error. Given no other standard, researchers must rely on human raters, but desire to quantify the inconsistency, accounting for the human error and validating the study.

To generate the standard against which to compare the data from the device, at least two raters will review a video of the subject during the test period and classify the activities using the same categories as the device. However, these ratings may be inconsistent, causing the standard against which the device is tested to contain an error, undermining the reliability study of the device. In order to quantify this error, we will select and implement a method of determining inter-rater reliability.

Background:
A popular and simple statistic to measure inter-rater reliability for nominal data is Cohen’s Kappa, which takes the joint probability of agreement of two raters and subtracts the probability of their agreement purely by chance, then scales it down by a factor proportional to the same probability of chance agreement (Cohen, J. 1960). A similar statistic for more than two raters is Fleiss’ Kappa, a generalization of Cohen’s Kappa (Fleiss, J.L. 1971). Within these statistics there are several ways to calculate the probability of chance agreement, including the sum of the product of the probabilities of each rater choosing the same activity, and the sum of the square means of the probabilities of each rater choosing the same activity.

Some, however, argue against the necessity of accounting for chance agreements, as it cannot accurately account for the certainty of a rater in rating each interval, and thus the degree to which the rater resorted to chance. As an alternative, versions of specificity and sensitivity are recommended, but without a gold standard, this is not easily attained (Uebersax, J.S. 1987). Depending on the choice of how to calculate the probability of chance agreement, however, the Kappa statistics may still be more appropriate than the raw joint probability of agreement.

Another alternative is to calculate the intra-class correlation, which can be approximated by weighted Kappa statistics (Fleiss, J.L. and Cohen, J. 1973). This method is appropriate for any number of raters. Since it is more useful for ordered data than nominal data (Fleiss, J.L. and Shrout, P.E. 1979),
this method may be useful in determining the inter-rater reliability of the duration of each activity, whereas Kappa might be more appropriate to determine the reliability of the raters’ choice of activity.

**Hypothesis:**

This study will calculate each of the statistics mentioned above to the ratings of each video. Assuming that the reliability of each rater does not change, the inter-rater reliability should also be constant. The consistency of the statistics between videos will therefore be examined to help determine the usefulness of each statistic in the reliability study. This may help conclude if accounting for chance agreements using a statistical approach is useful, as J. Cohen claimed (1960), or not, as J. Uebersax posited (1987). It may also demonstrate the accuracy of the approximation of the intra-class correlation by weighted Kappa statistics. The hypothesis is that accounting for chance agreements is useful but that its usefulness depends on the method of calculating the probability of a chance agreement, and that Kappa statistics and intra-class correlation will both be useful in measuring different aspects of reliability: Kappa may be more useful in determining the reliability of which activity is chosen in a given interval, and intra-class correlation may be more useful in determining the reliability of the duration of an activity.

**Methodology:**

To begin, a method of collecting and rating data independently of the device was required. This data will be compared to that of the device. One method used by other researchers was for annotators to be present at test sessions and record their observations using personal digital assistants (PDAs) (Parkka, J., 2006; Ermes, M., 2008). Under this method, a program would be written for the PDA to record the time at which a researcher selected each activity from a menu, and for this to be recorded as the start time of the selected activity and end of the previous activity. This method seemed simple and effective; however it had several drawbacks, the most significant of which follow. The presence of two or more trained researchers would be required at all test sessions, which would be less convenient and more intrusive. Only the minimum amount of data required would be collected. There would be no room for amending the annotations if definitions for a particular activity were amended due to early results.

The most favorable alternative found in literature was used by, among others, A. Shumway-Cook in her study of mobility in the elderly (2002). The subjects and their activities would be recorded using a video camera. This method requires only one researcher to accompany the subject, and raters would be able to review the recording at any subsequent time. Provided that the subject is kept in frame at all times, the video would be reliable and available for reconsideration if activities are redefined.

We selected the latter method, and after some investigation, selected the Sony DCR-SR68/R. In order to ensure full coverage of the subject, we added the accessories of a wide-angle lens and a high-capacity battery for longer recording times. Excluding any unforeseen hardware issues, the only remaining obstacle to the accurate acquisition of data is timing: the video should be synchronized with the data from the device. This issue has yet to be addressed, and its resolution will be included in the final report of this study.

The data to be rated will be collected during several test sessions in which a video recording will be taken of each subject wearing the device while performing several daily activities. There may be sessions in a controlled environment (i.e. the physical therapy building) while others will be conducted outside of the controlled environment. The details of these test sessions have been written by several physical therapy students under the guidance of G. Fulk, and submitted for review by the IRB. The recording will be performed by a trained researcher with a handheld video camera. Informed consent will be obtained from the subjects, and subjects’ confidentiality will be maintained.
A previous study conducted by E. Sazonov included the use of a virtual instrument generated in National Instruments’ LabVIEW environment to view and annotate a video. This same software could be modified for use in this study. The modified software should be able to display the video frame by frame, hold a rater’s annotations, and display the device’s output. This modification is already underway.

The video will be transferred to a computer and annotated independently by at least two raters using the software mentioned above. Each rater’s annotations will be output to a file for analysis. Once each rater has annotated the video, the reliability of the raters can be analyzed using the statistics described in the background section.

For each video, first the joint probability of agreement will be calculated as the number of agreements divided by the total number of agreements and disagreements. This by itself is a potentially useful statistic, and it is a factor in the computation of Kappa. Next, several different values of the probability of chance agreement will be calculated using various statistical methods. These will be combined with the joint probabilities of agreement to arrive at Kappa.

A program will be written in MATLAB to expedite this analysis. These statistics will then be compared and shown to agree or disagree with the hypothesis. Finally, the results of the analysis will be reported and interpreted to verify the reliability of the raters, and thus also the standard against which the device will be measured.

**Timeline:**
These activities are predicted to take the corresponding amounts of time:

- Writing Annotation Software: 4 weeks
- Data Collection and Annotation: 12 weeks
- Writing Analysis Software: 2 weeks
- Application and Writing Thesis: 18 weeks
References:


