**CS306: Data Analysis and Visualization**

**Lab 2: Report**

**Rajdeep Pinge 201401103**

**Aditya Joglekar 201401086**

**Objective: To use histograms and box plots as preliminary analysis tools for comparing two or more samples/populations**

Compare two methodologies that are widely used in subjective assessment of video quality.

1. Absolute category rating (ACR) in which the human observer is instructed to rate the quality of displayed video using adjectives: from ‘Bad’ (equivalent numerical score is 1) to ‘Excellent’ (equivalent numerical score 5).
2. Subjective Assessment Methodology for Video Quality (SAMVIQ) in which a continuous scale is used: 0 implies bad quality while 100 means excellent quality.

**Experiment1: Compare the quality scores from ACR and SAMVIQ methodologies by creating box plots**

We wrote a custom MATLAB function to generate the box plots. It is a simplistic box plot displaying the 5 measures of data ie Min, Max, Q1, Q2, Q3.

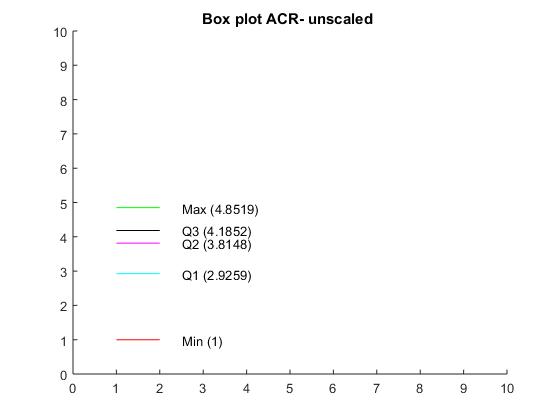
We have generated the following box plots:

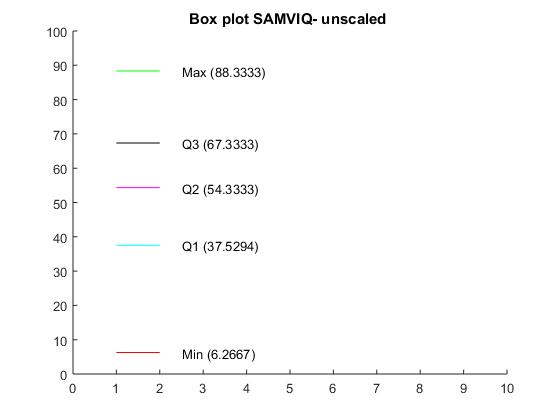
1. Unscaled: both datasets

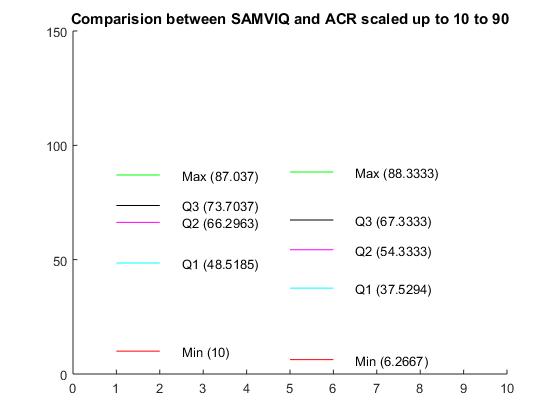
2. Scaled comparison between the two datasets, ASR scaled to 10-90 range

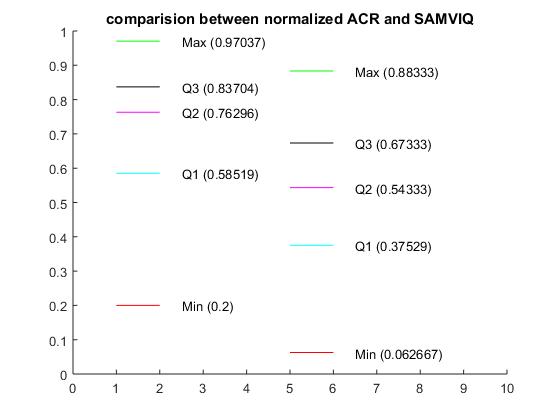
3. Scaled comparison between the two datasets, both datasets normalized to range [0, 1] by dividing by the maximum value in each dataset respectively.

The box plots are as follows:





****

****

**Question**

**Which of the two exhibits higher variability around the median? Justify your answer (hint: think of the qualitative differences between the two methods, and also the fact that different number of observers were used for each method)**

We use the interquartile range for this purpose

ACR\_Normalized:

Q3= 0.837 Q2= 0.5852 IRQ = 0.2518

SAMVIQ\_Normalized:

Q3= 0.6733 Q2= 0.3753 IRQ = 0.2980

SAMVIQ has slightly more variability around the median.

// insert values for the scaled up version

// also insert the standard deviation calculated, agrees with above as num outlier less

This is probably because of two reasons:

The SAMVIQ dataset values were averaged ratings on a continuous scale of 0 to 100. ACR values were averaged values on a discrete scale of 1-5. Thus as the domain of SAMVIQ data was larger, the possible values which the SAMVIQ dataset could contain was larger. Moreover, the number of voters were larger for the SAMVIQ dataset. Thus, the possibility of variability was larger in the second dataset.

**Which of the two, ACR or SAMVIQ, leads to mathematical outliers? Any reason(s)?**

**(assume points beyond Q3 + w\*IQR or Q1 - w\*IQR with w = 1.5, as mathematical outliers)**

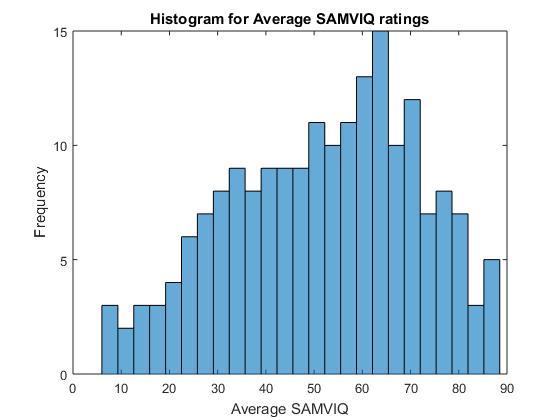
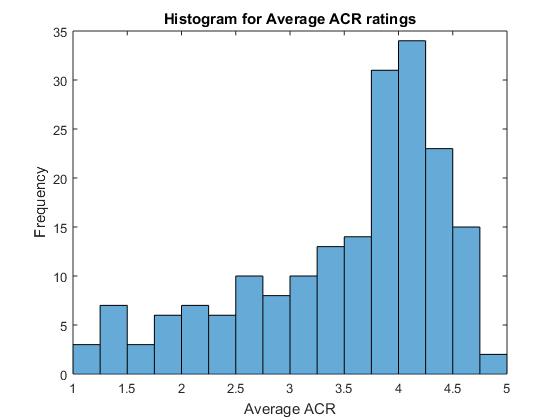
Number of outliers in ACR data: 2

Number of outliers in SAMVIQ data: 0

There are no outliers in the SAMVIQ dataset.

The inter-quartile range of ACR is lower than SAMVIQ. Hence, the fences for the outliers in ACR are closer to Q1 and Q3. Thus, the range of values which are considered acceptable (not outlier) is smaller for ACR. In other words the data in ACR is packed more tighly around the median than in SAMVIQ. Hence, it is easier for a point further away from the median to be a outlier.

**Plot the histograms for the two cases.**

****

**Question:**

**Describe the skewness of the two histograms.**

The SAMVIQ graph appears to be more symmetrically distributed. The ASR historgram on the other hand is more skewed towards the right. This, indicates that ratings in the histogram are positive biased ie a large number of ratings are on the higher side of the scale.

This could be because in the first case the ratings are discrete, so many ranges might get merged into one. For instance good and very good might be come under 4. In contrast the SAMVIQ graph has continuous metric which allow for a finer distinction in quality. So maybe most videos are good or better than that. In ASR these can’t be distinguished. But in SAMVIQ what would be in good category in ASR might be below the median, thus distributing the data more symmetrically.

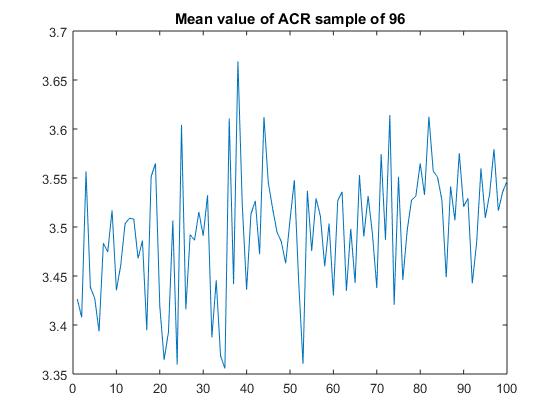
**Can we comment on skewness without plotting the histogram?**

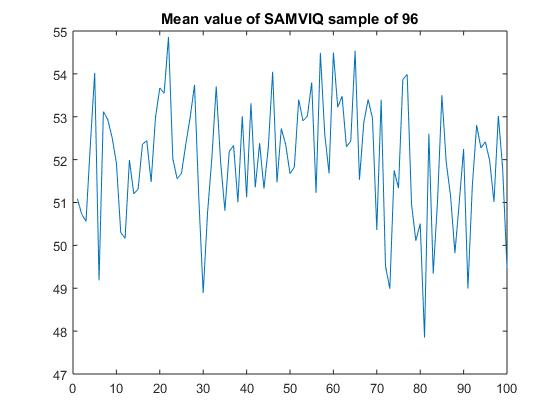
Yes, we can get an idea of skewness by looking at the box plots**.** In the ACR normalized box plot, the median is around 0.8 and the interquartile range is in upper half of the 0.0 – 1.0 scale. This indicates that around 50% of the data is concentrated in the upper 25% of the scale. Thus there is a bunching of data towards the upper half. This indicates a negative skew. This could be though of in terms of a kind of density. In a negative skewed distribution, median is at say 0.8. This means 50% data is in the 30% of the range. While rest is in the 70% range. ie concentration of data is more towards the right.

**Randomly sample half (i.e. 96) the scores for each methodology, and compute the mean in each case.**

Since, the sampling is random, we need to make sure that we do it enough number of times so as to ensure that an expected or average value of the random sample. If we do it only once, we might end up with a value which may not be very typical.

We computed the mean of the half sample over 100 iterations. The plots in both cases are as follows.





Mean for the entire dataset for both cases are as follows:

mean\_acr = 3.4975

mean\_samviq = 52.0036

**Question**

**How does it compare with the mean from the original data? Based on this, can we use only 96 samples (instead of 192) to compare ACR and SAMVIQ?**

Note that the random sampling also gives roughly the same mean as the complete dataset.

This can be readily seen from the above plots.