Quantile Metric: A New Approach to Compare Different Aggregation Methods for Point and Interval Estimates

Ying Han & David Budescu

Department of Psychology, Fordham University, New York

Abstract

The Quantile metric is a new standardized, easy-to-use tool that facilitates comparisons of forecasting performance of different aggregation methods, group sizes, elicitation methods and quality measures. The aggregated performance measure (e.g., Brier score, Q scores[1], hit rate[2], MAE[3], etc.) is mapped onto the cumulative distribution of individual performance scores of the same measures (cumulative distribution of individual Brier scores, Q scores, MAE, etc.) and the quality of the aggregated performance is evaluated by comparing it to the distribution of individual forecasters. We demonstrate the use of this new method with an empirical data set.

Methods

Data Set:

60 graduate students forecasted 40 target stock prices using point and 50%, 70% and 90% probability-interval estimates (Budescu & Du, 2007).

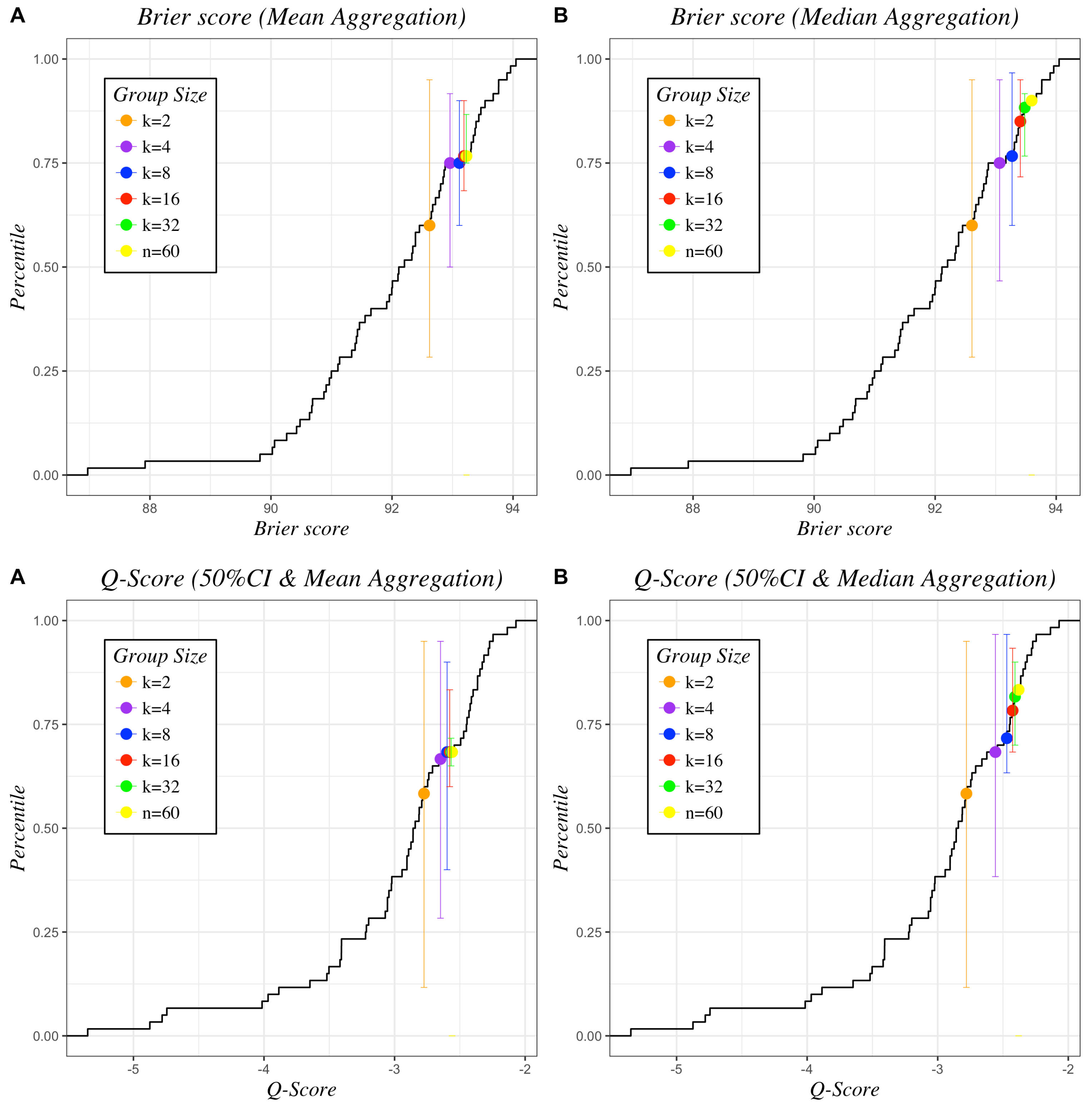
Procedures of Quantile Method:

1. Computed individual forecasting measures.
2. Obtained cumulative distributions for all individual forecasting performance measures (Brier scores and Q scores) across all the participants.
3. Randomly selected 32 (of the 60) participants and randomly assigned to smaller groups and analyzed as 16 groups of size k = 2 , 8 groups of size k = 4, 4 groups of size k = 8, 2 groups of size k = 16 and 1 group of size k = 32.
4. Computed aggregated group estimates in each group using mean and median aggregation and compute corresponding performance measures (Brier or Q scores).
5. Step 3 &4 were repeated 100 times.
6. For 100×32/k groups that have the same group size (k), we computed averaged aggregated group performance measures (mean and median aggregations) . We also obtained 90% empirical confidence interval for each averaged aggregated group performance measures based on 100 replications.
7. Aggregated results were mapped onto the corresponding individual cumulative distributions both numerically and graphically.

Questions/comments are welcome!
Contact: yhan23@fordham.edu

Results

Demonstration I : Comparison of different aggregation methods and aggregation group sizes



- Median aggregation in general yields higher forecasting performance compared to mean aggregation for all group sizes[4].
- Larger group size yields better aggregated results and the variation of the aggregates is reduced when the group size increases.
- The effect of group size is more salient in median aggregation.

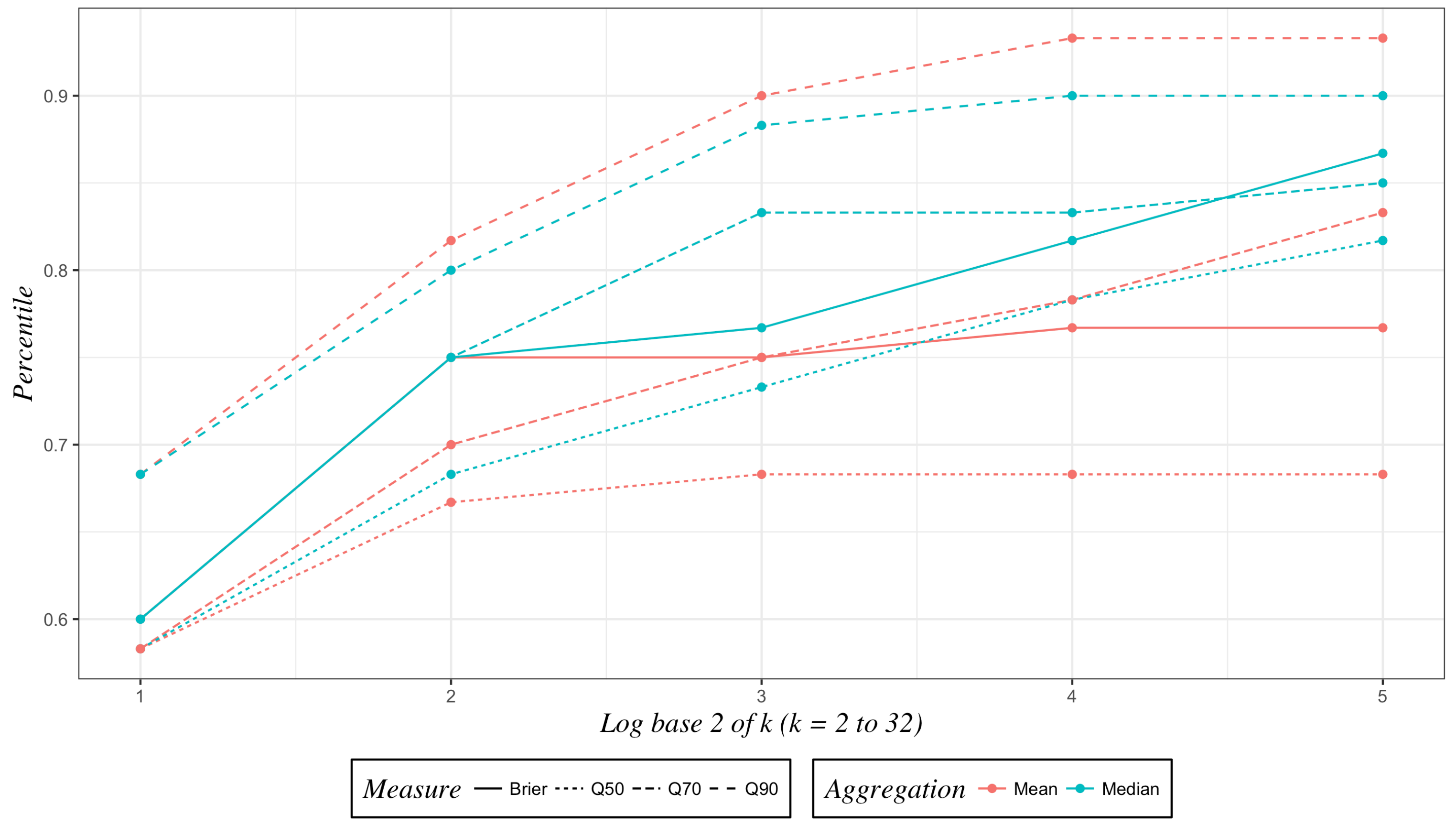
	k=2		k=4		k=8		k=16		k=32		n=60	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median	Mean	Median	Mean	Median
Brier	92.624	92.585	92.959	93.041	93.116	93.230	93.189	93.354	93.229	93.441	93.231	93.595
Quantile	**0.600**	**0.600**	**0.750**	**0.750**	**0.750**	**0.767**	**0.767**	**0.817**	**0.767**	**0.867**	**0.767**	**0.900**
Q50%	-2.774	-2.777	-2.651	-2.550	-2.599	-2.463	-2.579	-2.423	-2.568	-2.400	-2.559	-2.380
Quantile	**0.583**	**0.583**	**0.667**	**0.683**	**0.683**	**0.733**	**0.683**	**0.783**	**0.683**	**0.817**	**0.683**	**0.833**
Q70%	-2.040	-2.027	-1.948	-1.895	-1.898	-1.842	-1.866	-1.814	-1.848	-1.799	-1.832	-1.777
Quantile	**0.583**	**0.600**	**0.700**	**0.750**	**0.750**	**0.833**	**0.783**	**0.833**	**0.833**	**0.850**	**0.833**	**0.850**
Q90%	-0.956	-0.957	-0.867	-0.875	-0.824	-0.837	-0.804	-0.823	-0.796	-0.825	-0.792	-0.819
Quantile	**0.683**	**0.683**	**0.817**	**0.800**	**0.900**	**0.883**	**0.933**	**0.900**	**0.933**	**0.900**	**0.933**	**0.933**
Hit Rate 50%	-0.168	-0.170	-0.180	-0.161	-0.186	-0.165	-0.187	-0.167	-0.191	-0.167	-0.175	-0.150
Quantile	**0.550**	**0.550**	**0.467**	**0.550**	**0.467**	**0.550**	**0.467**	**0.550**	**0.467**	**0.550**	**0.533**	**0.650**
Hit Rate 70%	-0.109	-0.106	-0.091	-0.083	-0.084	-0.069	-0.076	-0.053	-0.076	-0.037	-0.075	-0.025
Quantile	**0.500**	**0.500**	**0.617**	**0.617**	**0.617**	**0.700**	**0.617**	**0.700**	**0.617**	**0.817**	**0.683**	**0.950**
Hit Rate 90%	-0.080	-0.080	-0.059	-0.053	-0.046	-0.035	-0.036	-0.023	-0.032	-0.016	-0.025	0.000
Quantile	**0.550**	**0.550**	**0.600**	**0.600**	**0.767**	**0.767**	**0.767**	**0.917**	**0.767**	**0.917**	**0.917**	**1.000**

How to interpret a quantile score?

Aggregated performance (k=32) of mean aggregation is as good as or better than 76.7% of individuals.

Demonstration II : Comparison of different elicitation methods

Brier Score, Q scores for 50%, 70% and 90% (Mean & Median Aggregation)



- Across all group sizes in both aggregation method, Q scores of 90% CI yields the highest forecasting quality and Q scores of 50% CI yields the lowest (Q scores of 70% CI and Brier scores lie in the middle).
- The relationship between Brier scores and Q scores 70% CI varies across different aggregation methods and different group sizes.

Summary

- These demonstrations showed the versatility of quantile metric that can be easily and efficiently applied to various circumstances. It also led to some meaningful findings about aggregated forecasting.
- Median aggregation was superior to mean aggregation for point-probabilities and probability-interval estimates.
- Comparison of forecasting quality of probability-interval estimates showed that forecasting performance is sensitive to the level of confidence.

[1] Q score is defined as $Q(L, U, x) = -(\alpha/2)(U-L) - \max\{L-x, 0\} - \max\{x-U, 0\}$, where L and U are lower and upper bound of α probability interval and x is true value of the estimated quantity (Jose & Winkler, 2009).

[2] Hit rate is defined as the proportion of intervals that contain the true value. [3] MAE: Mean Absolute Error.

[4] Unlike Brier score, Q score of 50%, 70% CI, Mean aggregation performs slightly better than median aggregation for Q score of 90%.