Confidence Intervals for Radio Ratings Estimators

Richard Griffiths[1]
[1]Arbitron, Inc., 9705 Patuxent Woods Drive, Columbia, MD 21046

Abstract

Arbitron's current method for forming confidence intervals for radio rating estimates is based on the normal distribution model. It is well-documented that with estimates of proportions from sample surveys the usual confidence interval method based on the normal distribution is generally lacking in some important situations. One of those situations is when the estimated proportions are very small (or very large) and especially when sample sizes are also small.

Because Arbitron ratings represent estimates of small proportions, we recently re-evaluated our confidence interval methodology. This methodology was first implemented in the early 1980's when radio listening was less fragmented – ratings were generally larger – and the methodology was less likely to fail.

This poster displays the results of an empirical study designed to compare the currently-used Wald confidence intervals to some alternatives, including Clopper-Pearson intervals. The poster also examines the practical implications of making a change.

Key Words: Confidence Intervals, Wald, Clopper-Pearson, Wilson Score, Coverage Probability, Radio Ratings, Empirical Study.

1. Background

To produce estimates of radio listening audiences in the United States, Arbitron divides the country into about 300 geographical areas called markets. Arbitron then conducts surveys of a sample of households in each market.

General findings from these surveys are that, in any given market, about eight to 15 percent of people are listening to the radio at any given time, on average. This eight to 15 percent is known as the Persons Using Radio (PUR) rating for the market. More specifically, Arbitron's estimate of the percent of people listening to a given radio station within a market during any given quarter-hour is called the Average Quarter Hour (AQH) Rating.

If a market has 30 radio stations – larger markets tend to have more, smaller markets fewer – then, on average, 0.3 to 0.5 percent of people are listening to a particular station during any given quarter-hour. So, AQH ratings – estimated proportions of people listening to a station during a given quarter-hour – represent small proportions.

As an example, in June 2009, the Washington, DC market PUR was estimated to be 8.7 percent. Forty-five radio stations were listed in the Arbitron report. Of these, only 10 had AQH ratings of at least 0.3 percent. Seventeen had ratings less than 0.1 percent.

The station with the most listeners – Station 1 in Figure 1 below – had a 0.8 percent AQH Rating. This means that, on average, during any 15-minute interval, 0.8 percent of the people in the DC listening area were tuned to this station.

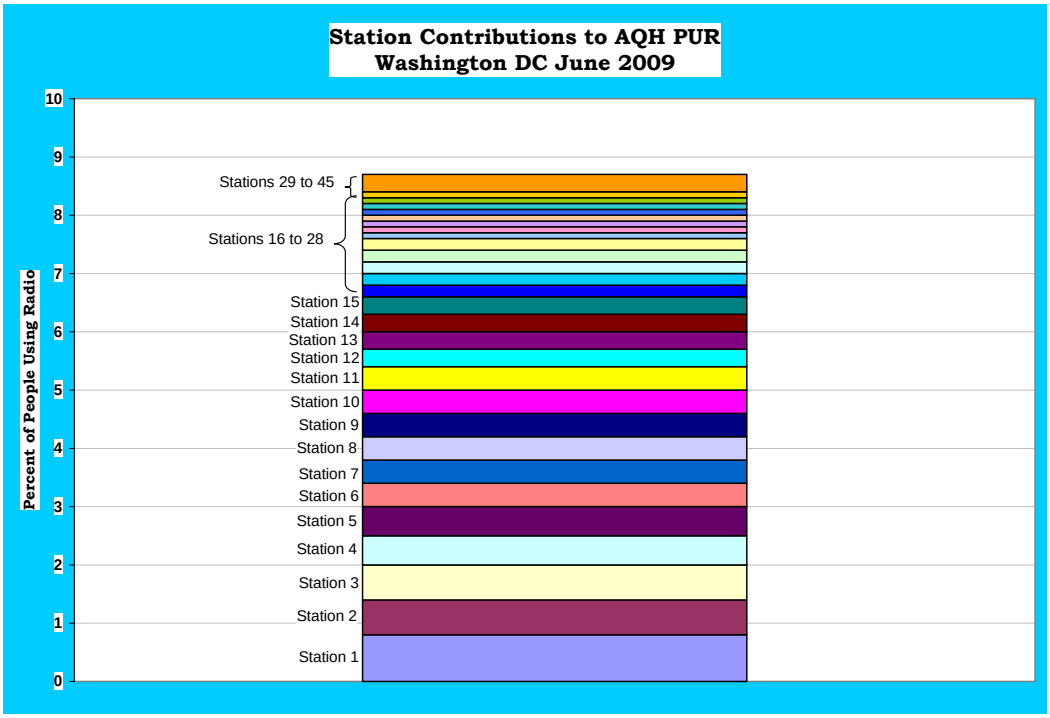

Figure 1. Breakdown of average quarter hour listening in the Washington, DC market in June 2009

2. The Problem

Arbitron's current confidence interval method for AQH Ratings is the Wald, normal-based approximation method:

$$\hat{p} \pm z_{1-\frac{\alpha}{2}} \cdot s.e.(\hat{p}),$$

where $\hat{p}$ represents the station AQH rating.

The primary problems with the Wald confidence intervals for AQH ratings are twofold:

- They can have low coverage probabilities for small proportions and small sample sizes.
- Wald confidence intervals can give nonsensical limits:
  - Lower limits that are less than zero.
  - Upper limits that are greater than one.

(See Wikipedia Contributors, 2009; Cochran, 1977, Chapter 3; Agresti and Coull, 1998.)

3. Proposed Solution

To address the concerns with the Wald intervals, we investigated two confidence interval methods without these shortcomings: the Clopper-Pearson and Wilson Score methods.

Clopper-Pearson method

The confidence limits of the Clopper-Pearson (CP) method are given by the following:

Lower Limit: $Beta\left(\frac{\alpha}{2}, x, n-x+1\right)$

Upper Limit: $Beta\left(1-\frac{\alpha}{2}, x+1, n-x\right)$

where x is the number of "successes" (listeners) and n is the number of "trials" (respondents).

This method was originally discussed in Clopper and Pearson, 1934.

Wilson Score method

The bounds of the Wilson Score method are given by

$$\frac{\hat{p}+z^2_{1-\frac{\alpha}{2}}/2n}{1+z^2_{1-\frac{\alpha}{2}}/n} \pm z_{1-\frac{\alpha}{2}} \cdot \frac{\sqrt{\left[\hat{p}(1-\hat{p})+\frac{1}{4n}z^2_{1-\frac{\alpha}{2}}\right]/n}}{1+z^2_{1-\frac{\alpha}{2}}/n}$$

This method was originally discussed in Wilson, 1927.

For use in calculating confidence intervals for Arbitron's AQH rating estimates, we used the ad-hoc procedure of substituting the effective sample size for n in the formulas to account for complex sample design.

4. Some General Facts About the CI Methods and Coverage Probabilities

Heading into our investigation, we were mindful of some general facts about the different confidence interval methods:

- The CP method tends to be conservative: coverage probabilities are generally larger than nominal confidence levels.
- The Score method is less conservative: coverage probabilities can be smaller than or larger than nominal confidence levels, but tend to be closer than the CP method.

- The Wald method tends to have poor coverage probabilities – less than nominal levels – for small/large proportions and small *n*.

5. Investigation Method

We conducted an empirical study of the confidence interval methods using Arbitron radio ratings data.

We used the following data:

- AQH ratings estimates for radio stations in Los Angeles, Chicago, and Houston markets.
- These estimates were from several months of sample surveys from 2007 and 2008.
- The estimates were for various demographic subgroups and parts of the day (dayparts).

The methods used in the investigation can be outlined as follows:

- Generate hundreds of sub-samples from the market full samples.
- For each sub-sample, calculate AQH ratings estimates for each station, by demo and daypart.
- Estimate/model the variance and effective sample size for each sub-sample estimate.
- Construct confidence interval limits – Wald, CP, and Score – for each sub-sample estimate.
- Examine proportion of intervals that include the full-sample rating (the assumed population proportion). These proportions serve as our empirical coverage probabilities.

6. Empirical Study Results

Summarizing the empirical coverage probabilities from the empirical study over all demographic groups, dayparts, and stations, we found the following:

- The CP method overshoots the nominal level by a little more than two percent.
- The Score method comes in slightly under nominal level.
- The Wald method is somewhat further under.
- Generally, though, none are far from the 90 percent nominal level.

See Figure 2 on the next page.

Looking at the results of the empirical study coverage probabilities by demo, we find that more pronounced differences start to emerge:

- For demographic groups with smaller sample sizes – Children aged 6-11, Teens 12-17, Males and Females 18-34 – the Wald method undershoots the nominal level by more: five to seven percent, on average.
- The CP method continues to slightly overshoot the nominal level, as expected.

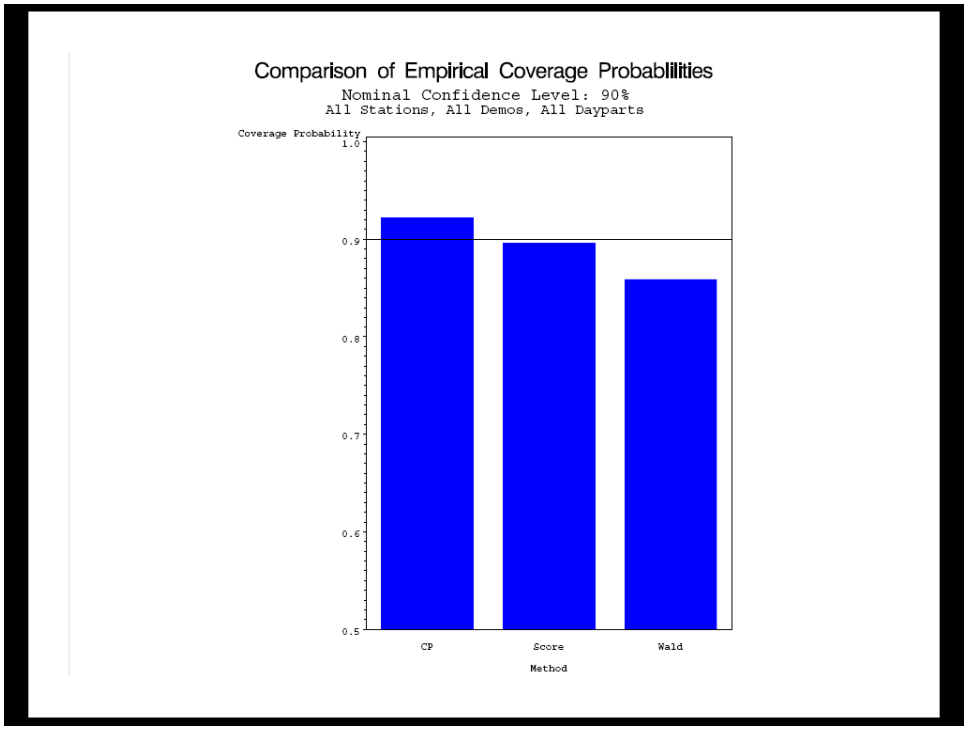

Figure 2. Empirical Coverage Probabilities, All Demos, Dayparts, and Stations.

- The Score method is sometimes over and sometimes under, but never by more than two percent.

See Figure 3.

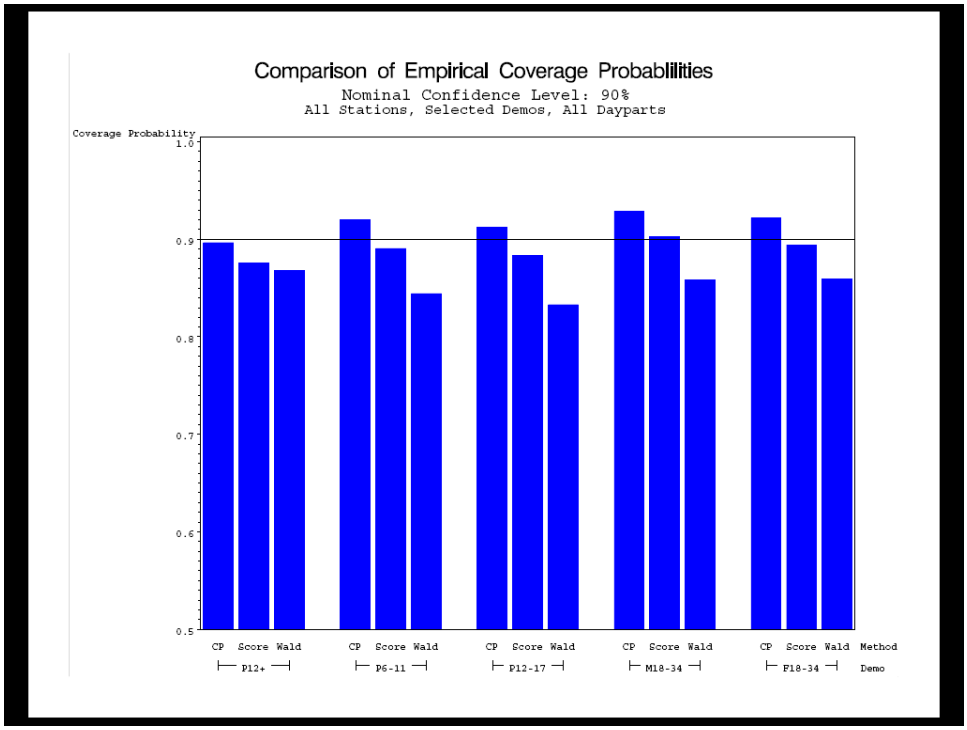

Figure 3. Empirical Coverage Probabilities by Demo.

Examining the results by the size of the sample on which the estimates are based, we found the following (Figure 4):

- As expected, the Wald method's empirical coverage probabilities are further from the nominal level, the smaller the sample size gets.
- They are close to 10 percent less than the nominal level for the smallest sample sizes, which in this study are between 75 and 125.
- As the sample sizes get larger, the Wald method comes closer to the nominal level.
- The CP method has somewhat large empirical coverage probabilities for all sample sizes, about one to two percent over the nominal level.
- The Score method's empirical coverage probabilities tend to be slightly under nominal level.

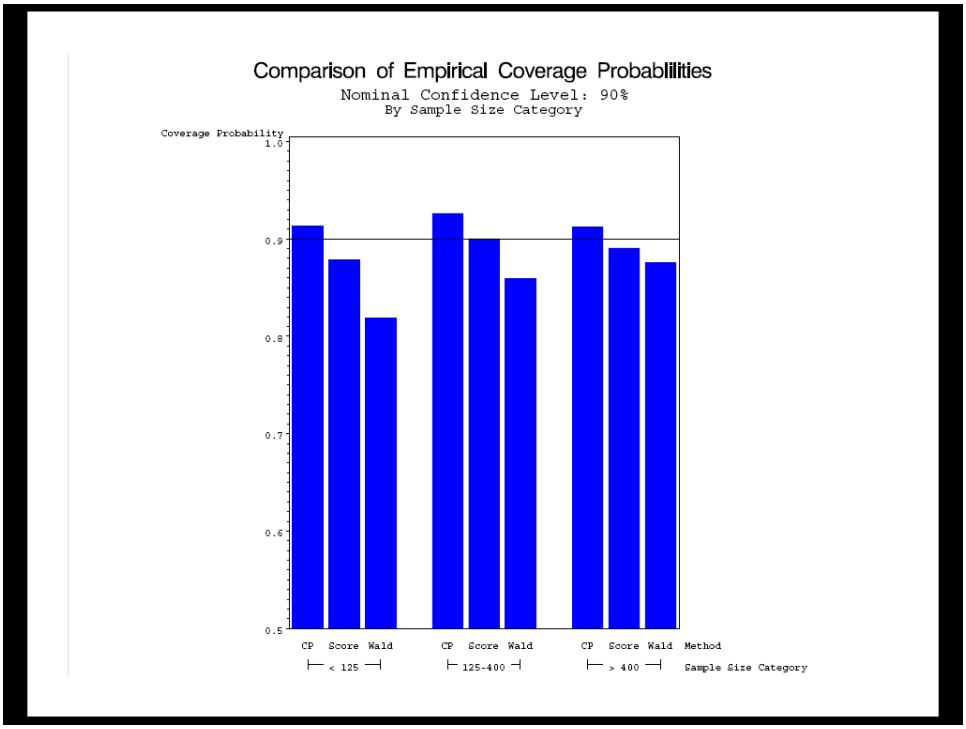

Figure 4. Empirical Coverage Probabilities by Sample Size.

Examining the results by sample size for the stations with the smallest rating estimates – less than 0.1 percent rating estimates – we found the following (Figure 5):

- The Wald method undershoots the nominal level by slightly more, on average.
- For the smallest sample sizes and smallest station ratings, the Wald method comes in about 13 percent under the 90 percent nominal level.
- The Score method comes in at about five percent under the nominal level for the smallest sample sizes and smallest station ratings.
- Otherwise, for even the smallest-rated stations, the Score method empirical coverage probabilities are close to the nominal level.
- The CP method is again generally conservative, by about three or four percent for the smallest-rated stations and medium to large sample sizes.

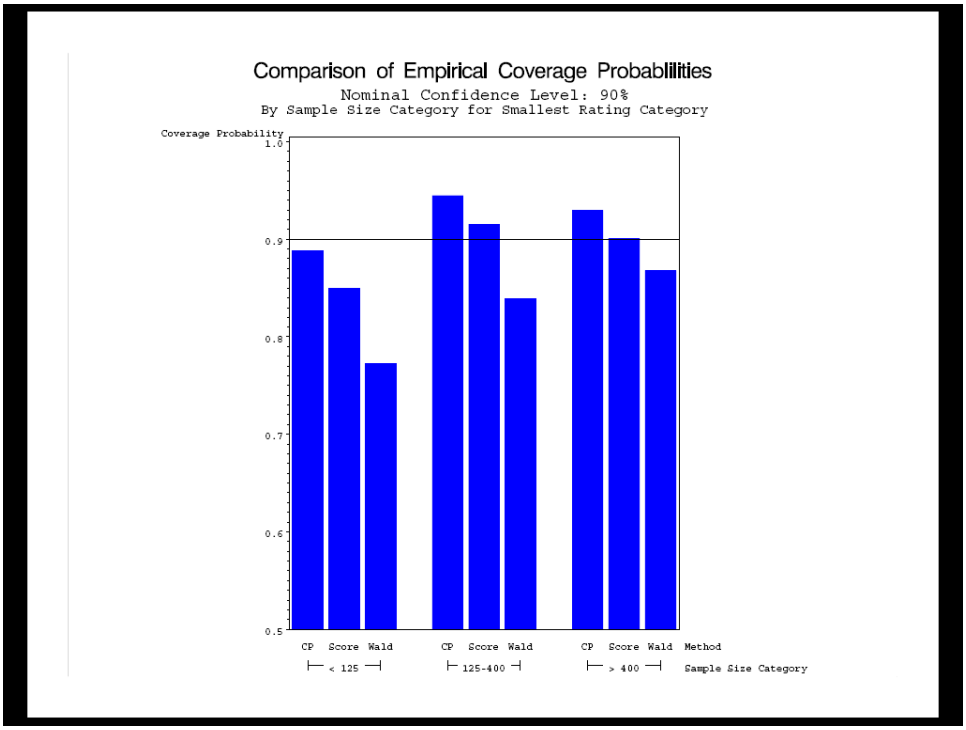

Figure 5. Empirical Coverage Probabilities for Stations with Small Rating Estimates.

7. Discussion

Our general expectations held up during the empirical study:

- The CP method tended to be conservative.
- The Wald method had problems with small sample sizes and small proportions (ratings).
- The Score method had empirical coverage probabilities that were sometimes less than the nominal, sometimes more, but generally close.

Perhaps a mild surprise was that the Wald method held up as well as it did. Almost, all station rating estimates are under one percent – definitely small proportions – yet it wasn't until the sample sizes dropped to around 100 that we encountered seriously deficient coverage probabilities with the Wald method. However, we still don't like the possibility of nonsensical confidence limits – negative lower limits. This and the fact that the Score and CP methods had at least slightly better empirical coverage probabilities, in general, reinforce our conviction to move away from the Wald method.

The choice between the CP and Score methods is more difficult. On the one hand, the general negative of the CP method is that it tends to be too conservative. However, in our empirical study, it didn't come out that conservative. Part of this is, no doubt, due to the sample sizes and, in particular, the effective sample sizes. None of the sample sizes were less than 75 in this study and the smallest effective sample sizes tend to be close to 500. (See Appendix A on **e**ffective sample sizes.)

The Score method also performed well. Its empirical coverage probabilities, while sometimes less than the nominal level, were generally closer to the nominal level than the CP method.

The fact that the Score method empirical coverage probabilities were five percent less than the nominal level for the smallest sample sizes and ratings is a little concerning. Given the microscopic nature of some station ratings, we wonder if these estimated proportions aren't in a region for which the Score method will have low coverage probabilities. (See Agresti and Coull, 1998, p. 122.)

Another consideration is that the confidence interval method needs to be applied to another type of Arbitron estimate, cume ratings. Cume ratings tend to be larger than AQH ratings, but have significantly smaller effective sample sizes.

While we tend to have a preference for the Score method, based largely on the literature and its performance in this empirical study, we are also aware of the need to further evaluate it, and the other methods, in some of the more extreme cases and for cume ratings.

8. References and Further Reading

Agresti, A. and B.A. Coull, 1998, "Approximate is Better than 'Exact' for Interval Estimation of Binomial Proportions," The American Statistician, 52, 2, 119-126.

Brown, L.D., T.T. Cai, and A. DasGupta, 2001, "Interval Estimation for a Binomial Proportion," Statistical Science, 16, 2, 101-133.

Clopper, C.J. and E.S. Pearson, 1934, "The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial," Biometrika, 26, 4, 404-413.

Cochran, W.S., 1977, *Sampling Techniques*, Wiley.

Kott, P.S., P. G. Andersson, and O. Nerman, 2001, "Two-Sided Coverage Intervals for Small Proportions Based on Survey Data," Proceedings of the 2001 Federal Committee on Statistical Methodology Conference.

Liu, Y.K. and P.S. Kott, 2007, "Evaluating Alternative One-Sided Coverage Intervals for an Extreme Binomial Proportion," 2007 Statistics of Income Paper Series, Internal Revenue Service.

Wikipedia Contributors, 2009, "Binomial proportion confidence interval," http://en.wikipedia.org/wiki/Binomial_proportion_confidence_interval, accessed August 3, 2009.

Wilson, E. B., 1927, "Probable Inference, the Law of Succession, and Statistical Inference," Journal of the American Statistical Association, 22, 158, 209-212.

Appendix A: Effective Sample Sizes

In our empirical study, we use the effective sample size in place of *n* in the confidence interval formulas. This is done to account for the complex nature of the sample design that the Arbitron radio ratings are based on.

The sample design can be briefly described as a stratified, cluster sample of households. All persons, aged six and older, are asked to participate.

There is some mild geographic over- and under-sampling applied.

The primary factor of the design that affects AQH rating effective sample sizes is a "repeated measures" factor:

- AQH rating estimates are averages of the estimated proportion of persons listening to a radio station during any given quarter-hour.
- Since Arbitron credits a person with listening or no listening in quarter-hour intervals, we effectively have many observations of a respondent's listening over a given time period. For example, for a rating based on the "morning drive" daypart – Monday through Friday from 6am to 10am – we effectively have 80 (=4 x 4 x 5; 4 quarter hours per hour, 4 hours in each day of the daypart, and 5 days) individual measurements for each respondent going into the rating estimate.

This "repeated measures" factor makes AQH rating effective sample sizes many times larger than the actual number of respondents in the sample.