ARF / ESOMAR
Worldwide Electronic and Broadcast Audience Research Symposium

San Francisco, USA
21st-24th April 1996

The Variability of Audience Measurement Data

and How to Live With it

by Tony Twyman and Steve Wilcox

The paper reviews the concept of sampling error applied to various methods of sampling and television panels in particular.

Approaches are described for calculating sampling errors in relation to:

Programme/Commercial Break Ratings
Commercial Impacts/Average Hours of Viewing
Channel Reach
Channel Share
Schedule Reach and Frequency
Changes Over Time

An approach to the use of sampling error is suggested given that audience measurement figures **have to** be used in making decisions even when the results lie within statistical confidence limits.

Panel analyses which can throw light on the contribution of sampling error are also discussed.

An awareness of broad levels of sampling error are essential to users and should play a part in designing research and fixing sample sizes.

Introduction

In broadcast media research, audience sizes are used as 'currency' in a trading system which treats audience research data as hard facts, not as estimates with a variable degree of distance from the (unknowable) truth. In parallel, programme producers regard every twitch of the audience figures as evidence of success or failure. These modes of usage involve a suspension of disbelief which collapses when users suffer from variability in audience estimates which are both unpredictable and/or contrary to expectation.

The collection of television audience measurement data is a highly intensive operation; via peoplemeters, the 24 hour second-by-second electronic monitoring of an aspect of human behaviour which often occupies most time after working and sleeping (in some cases more than those). Consequently it is very costly, and economic pressures operate to minimise sample sizes. This leads to variable data used as if they were invariable facts. When variability has to be faced it is then often referred to in terms of sampling error, which is regarded as a terminal illness from which some data may suffer rather than an ever-present extreme boundary within which normally, much smaller variations may occur.

When sampling error is invoked it is often used to prove that some unpopular statistic is not significant. The approach is usually based upon a simple binomial calculation of the 95% limits around a percentage and breaks down when statistics other than single percentages are under consideration, as they usually are. When users consult a theoretical statistician they may be told that the sampling error of anything other than a random probability sample cannot be calculated. Whilst this is true in terms of sampling error about the absolute, unknown, truth, virtually no media research meets these requirements. And yet it is essential that we have some idea about the variability of the data upon which so much depends based on our necessarily imperfect sampling methods and the desperately small samples which we have to use.

The form of the paper is therefore:

- Discussion and justification of the calculations of expected variability for the kind of samples which are used in broadcasting research.

- A short discussion of panel bias.

- Review of the calculation of sampling errors for the kinds of statistic used in decision-making in the media. Examples are given based on services carrying advertising in the UK. This is partly because of the greater complexity of uses associated with advertising and partly also because of the availability of a convenient database.

- A brief account of some work on UK radio research samples.

- A discussion of strategies for using 'sampling errors' in practice and approaches to investigating some of the components of sampling error in panels.

- Final conclusions and ideas for further investigation.

1. The Calculation of Sampling Error

If our TV or Radio audience measurement systems produce a rating of 20%, then we want to be able to measure our confidence in this estimate. Ideally we want to know how far this rating estimate could be from the "truth", where "truth" is defined to be the rating amongst the **whole population** that our sample or panel purports to represent. In this respect the required advice from the statistician is the size of the confidence interval, ie. the value of x where the user can be 95% confident that the "true" rating lies somewhere in the interval 20% plus or minus x. The value of x is a multiple of the sampling error.

Whilst the concept is simple, real life is not so straight forward. First of all, we need to be clear about the differences between what we would like to measure and what we can measure when calculating sampling error. In this respect we must consider the potential effects of sample selection bias and non-response bias.

Given an adequate sampling frame, selection bias should not exist with probability sampling. It is an issue with quota sampling where the **interviewer** selects respondents, albeit to a set of rules which attempt to minimise the selection bias. Non-response affects both quota and random samples. Within a quota sample, it is usually impossible to separate selection and non-response bias.

But having attempted to differentiate between quota and random samples, we must ask ourselves how different they are in practice. In most random samples, a proportion of the non-response element is attributed to non-contact (e.g. no answer at an address after a number of calls) and it is difficult to see how this differs from the selection bias in quota sampling!

Whilst most audience measurement surveys employ stratification (explored later) in an attempt to reduce potential bias, we cannot normally measure how large this bias is. Therefore we must acknowledge that it is there without knowing how large it is and try to minimise it.

The remaining component of the sampling error is probably most correctly referred to as the precision of the sample, but there will be more consistency with other work if we refer to this as the sampling error. In the context of the confidence interval, this means that we will be defining the "truth" to be the actual rating amongst the population that our sample actually manages to represent.

When a new panel is selected, sampling error can by chance create a panel which is slightly different from the 'sampling method population'. This difference then becomes a temporary panel 'bias'. As panel members leave and new members join the panel, this initial 'bias' dissipates, probably to be replaced by a new panel bias. Working within an unchanging panel of individuals, the sampling variability arises from the behaviour of those individuals and corresponds to the concept of precision metnioned above. As comparisons are made over time, changes in 'bias' are also involved, creating a total sampling error which increases as the panel changes. In panel operations where enforced turnover is employed, this increase in sampling error is accelerated.

So what is this sampling error? Well, in order to explain it, lets put ourselves in the hypothetical - but very fortunate - position of being able to draw a large number of independent samples, all to the same design and designed to measure the same thing. It is natural to expect that a large number of the samples will result in a measurement of the rating which is close to the (biased) "true" rating but that a few will be either a long way above or below the "true" rating. The distribution of these ratings estimates about the "true" rating follows the familiar "bell-shaped" curve of the Normal probability distribution. The peak of this curve would coincide with the "true" rating.

The sampling error for just **one** of these samples, is the "root mean square deviation from the mean". So if we have N independent samples producing a rating of R_i for sample i, the sampling error for **one** sample is:

$$\sqrt{\frac{1}{N}\sum(R_i - \bar{R})^2}$$

The larger the number of samples, the closer $\bar{R}$ (average rating of the N samples) will be to the "true" rating.

Then assuming that the distribution of the ratings estimates does follow a Normal probability distribution, we would expect:

> 68% of samples to give a rating estimate within plus or minus 1 sampling error of the "true" rating.
>
> 90% of samples to give a rating estimate within plus or minus 1.65 sampling errors of the "true" rating.
>
> 95% of samples to give a rating estimate within plus or minus 2 sampling errors of the "true" rating.

and so on. And we can turn this around to say that for **one** such sample, we would be 68% confident that the "true" rating is within plus or minus one sampling error of our sample estimate.

Now, if the size of each sample is increased, there will be less dispersion about the "true" rating: the sampling error will be smaller. In fact the sampling error will decrease in inverse proportion to the increase in the square root of the sample size. So if the sample is four times as large, the sampling error will be halved. And this principle enables us to calculate the sampling error using only **one** sample of n individuals. The trick is to assume that we have n independent samples each comprising only one individual.

Then if each individual i has a rating of r_i (where r_i is either zero or 100%) and the sample average is $\bar{r}$, then the sampling error for **each individual** is:

$$\sqrt{\frac{1}{n}\sum(r_i - \bar{r})^2}$$

Because we are now talking about an individual, we call this the "sample standard deviation". Now, we actually have n samples each of one individual, so the sampling error for the sample (average) rating $\bar{r}$ is:

$$\text{s.e.}(\bar{r}) = \frac{\text{Sample Standard Deviation}}{\sqrt{n}}$$

As explained above, this allows us to calculate the confidence intervals around $\bar{r}$ - the sample (average) rating - within which we expect the "true" rating to fall with different levels of certainty.

Example

Suppose we have a sample of 1000 individuals and an (average) rating of 20%. This means that 200 individuals have a rating of 100% (r_i =100) and 800 individuals have a rating of 0% (r_i = 0). The sample standard deviation is:

$$\sqrt{\frac{1}{1000}\{200x(100-20)^2 + 800(0-20)^2\}} = 40$$

And to get the sampling error, we have to divide by the square root of the sample size, so:

$$\text{Sampling Error} = \frac{40}{\sqrt{1000}} = 1.3$$

It is no coincidence that the more familiar formula for the sampling error of a percentage p:

$$\text{s.e.}(p) = \sqrt{\frac{pq}{n}}$$

delivers the same answer. ie. if p = 20% and q = 100-20% = 80%, then:

$$\text{s.e.}(p) = \sqrt{\frac{20x80}{1000}} = 1.3$$

This is because in a yes/no (e.g. viewed vs. didn't view) situation, the two formulae for the sampling error are mathematically the same. This doesn't apply to other measurements of viewing such as average hours per day.

2. Panel Bias

All samples have a potential for bias. Even a probability sample with a high response rate is likely to miss people who are very busy and most surveys are probably avoided by people with low literacy. With television panels, even where selection by the interviewer is avoided, there are the additional problems of cooperation bias. Response rates are traditionally low in panel research. To compensate for this, panels are subject to a high degree of stratification, controls and weighting. This results in the weighted sample matching universe profiles on most classifiers known to influence viewing. This high degree of control may well increase precision, i.e. reduce variation, but still represent a population biased relative to the total population. It is possible to have a sample with a high level of bias but low variability. This really means high absolute sampling error but low variability making it possible to measure differences quite precisely. An example of this is hall testing of products where due to the homogeneity of respondents co-operating in hall tests and the small number of interviewers involved, high degrees of precision were found but with a great potential for bias. There is some analogy here with television panels and it is essential to try to minimise bias. Within the classifiers by which television panels are controlled there may be other characteristics which are correlated both to viewing behaviour and readiness to join a panel. One example, identified in the UK many years ago, was 'claimed weight of viewing' where it was shown that viewing claims did correlate positively with actual viewing levels and that lighter viewers were less likely to join and more likely to drop out of television panels. This has been adopted as a rather awkward panel control in the UK. Periodically, BARB's contractors have conducted studies which take all the known characteristics of panel members and relate them to actual viewing levels via analyses of variance to establish the most important determinants of viewing. Claimed weight of viewing may be regarded as a lifestyle indicator and it is possible that more indirect questions about amount of leisure and interest in television may be alternative approaches.

Probably within advertising there is more interest in the precision of television audience data and its capacity to measure differences and trends, than there is concern about unknown bias. The public service broadcaster has most interest in bias in relation to proving that all sections of the public are served.

It is argued here that bias should be of concern to all. The suggested strategy for dealing with it is to measure characteristics on the television panels which are correlated to viewing behaviour and check the profiles of the panels against the universe on these characteristics. If there is a significant mismatch then new stratifications and controls are needed. Despite a history of such investigations, apart from claimed weight of viewing, BARB has not found any evidence of panel bias in the UK.

3. The Calculation of More Realistic Sampling Errors

The example in the previous section is a very simplistic calculation of sampling error which assumes that we have a simple random sample. It takes no account of the things we do to constrain the cost of the research, to improve its usefulness or to make it more representative. However, the basic sampling error formula still holds, in that:

$$\text{Sampling Error} = \frac{\text{Standard Deviation}}{\sqrt{\text{SampleSize}}}$$

The sample design elements which can affect sampling error are:

Stratification - through sample/panel control and weighting, stratification is designed to improve representativeness and can improve precision.

Disproportionate sampling - the UK TV panel is disproportionately sampled by geography and demographics.

Weighting - partly to re-profile disproportionate sample design elements but also to control other key profiles to universe targets.

Clustering - both the TV panel and the Radio measurement sample cluster individuals within households; the Radio sample also clusters households within sampling points.

Each of these factors can have an effect on the standard deviation and/or the effective sample size. We should really re-write the formula to be:

$$\text{Sampling Error} = \frac{\text{Standard Deviation}}{\sqrt{\text{EffectiveSampleSize}}}$$

The standard deviations and effective sample sizes must be calculated from the actual sample data. Using formulae with different levels of sophistication we can measure the effect of each design element on the sampling error.

Finally we must recognise that there will be a different standard deviation for each audience measurement (e.g. spot ratings, daily average hours, weekly average hours, schedule reach and frequency), for each channel and for each demographic reporting category. Hopefully it will be possible to model the variations in these standard deviations so that sampling errors are more readily accessible.

In this paper we have concentrated on the requirements for advertising sales, covering the following key audience measurements:

Average Hours of Viewing (accumulated commercial impacts)
Channel Share

Channel Reach
Commercial Break Ratings (Programme Ratings)
Schedule Reach and Frequency

In all cases we recognise that change over time is usually required, therefore sampling error on change has also been covered.

The sampling error examples which follow are based upon the UK BARB TV panel using the 3300 homes in England (ie. excluding the 1200 homes in Wales, Ulster and Scotland). This panel is sampled disproportionately across 12 regions and within each region is sampled disproportionately by demographics. The viewing data is taken from March 1994, March 1995 and April 1995.

4. Average Hours of Viewing

Average hours of viewing are highly correlated with accumulated commercial impacts and are undoubtedly a key indicator of performance for each broadcaster.

The first example is daily average hours of viewing to ITV by all adults on a single day in March 1995. Table 4.1 shows the effects of each sample design criteria on the components of sampling error - the effective sample size and the standard deviation - and on the resulting 95% confidence interval.

Table 4.1 All Adults ITV Single Day

Average Hours = 1.43

Complexity	**Effective Sample**	**Standard Deviation**	**Confidence Interval**
Unweighted	6872 *100*	124% *100*	±3.0% *100*
Stratification	6872 *100*	124% *100*	±3.0% *100*
Weighting	4743 *69*	122% *98*	±3.6% *119*
Clustering	2811 *41*	104% *84*	±3.9% *131*
Full Specification	2185 *32*	106% *85*	±4.6% *152*

Using the most naive calculation of sampling error (ie. assuming that the panel is a simple random sample with no weighting), the effective sample is the same as the actual sample of 6872 adults. The standard deviation is 124% of the average hours and the resulting 95% confidence interval is 1.43 hours plus or minus 3.0%. This defines a base against which we can measure the individual and combined effects of stratification, clustering and weighting.

To illustrate the effects of stratification, we have used the split between those individuals who can or can't receive satellite channels. If we use panel control and weighting to ensure that we have the correct proportion of homes with satellite, then we can legitimately reduce the standard deviation - there is no change in the effective sample size. Stratification effects are often much smaller than you would expect and have no effect in this example.

The BARB TV panel is weighted to re-profile the disproportionate structure and to control the representation of key determinants of viewing behaviour and reporting categories. The sample is degraded by the extent to which the range of weights introduces extra variability, resulting in an effective sample size which is only 69% of its original value. The weighted standard deviation can go up or down depending upon which individuals get the larger weights - in this case we have a marginal decrease. The resulting confidence interval is 19% wider.

Clustering recognises that the viewing of individuals in the same household will be correlated to some extent. The basic sampling unit becomes the household, the hours of viewing for that household is the average of all the adults in the household. In the calculation of overall average hours, each household is effectively given a weight which is the number of adults in the household. This results in a dramatic reduction in the effective sample size to only 41% of its original value. This is partly offset by a reduction in the standard deviation to 84% of the base line value. The resulting confidence interval is 31% wider due to clustering.

Finally the combined effects of stratification, clustering and weighting are shown in the results which take account of the full sample design specification. The effective sample is reduced to 32% of its original value. The standard deviation is reduced to 85% of the base line value. The net result is a confidence interval which is 52% wider than that calculated using the most naive assumptions.

A very different example is hours of viewing by 16-24's to all satellite channels, again for a single day. This involves a small sub-group and a restricted availability situation in that only 20% of the population have satellite. The average daily hours of viewing is only 0.2 hours per person. The results are shown in Table 4.2.

Table 4.2 16-24 Adults Total Satellite Single Day

Average Hours = 0.20

Complexity	**Effective Sample**	**Standard Deviation**	**Confidence Interval**
Unweighted	927 *100*	384 *100*	±25% *100*
Stratification	927 *100*	348 *91*	±23% *91*
Weighting	719 *78*	369 *96*	±28% *109*
Clustering	591 *64*	363 *94*	±30% *118*
Full Specification	475 *51*	314 *82*	±29% *114*

This time, the most naive calculation of sampling error has an effective sample size of 927 and a standard deviation of 384%. This smaller sample and greater variability combine to generate a confidence interval of 0.20 hours plus or minus 25%.

Because the stratification is so directly related to the measurement, there is a substantial 9% decrease in the width of the confidence interval.

Weighting and clustering have much less effect upon the effective sample, standard deviation and resulting confidence interval. This is because there is no disproportionate sampling amongst 16-24's and there are fewer 16-24's per household.
Again the combined effects of stratification, weighting and clustering are shown in the results which take account of the full sample design specification. The effective sample is reduced to 51% of its original value. The standard deviation is reduced to 82% of the base line value. The net result is a confidence interval which is only 14% wider than that calculated using the most naive assumptions.

Over longer periods of time, individuals tend to get more similar in terms of their cumulated hours of viewing resulting in a smaller standard deviation. Therefore sampling errors and confidence intervals get smaller, as shown in Table 43.

Table 4.3 Average Hours for Longer Periods

Confidence Intervals

	All Adults ITV	**16-24 Adults Total Satellite**
1 day	±4.6% *100*	±29% *100*
1 week	±3.7% *82*	±20% *69*
4 weeks	±3.4% *75*	±17% *57*

In these examples, the majority of the reduction occurs between 1 day and 1 week. This reduction is larger for 16-24 adults total satellite - again we are seeing a relative benefit for the minority measurement, this time gained by averaging over a longer period.

Now staying with four week average hours, the data in table 4.4 is designed to explore the relationships between demographics and between channels.

Table 4.4 4 Week Average Hours

	Sample Size	Effective Sample	Standard Deviation	Average Hours	Confidence Interval
Total TV					
All Individuals	8529	2120	56%	95	±2.4%
Adults	6872	2185	56%	101	±2.4%
Housewives	3313	2227	59%	115	±2.5%
Men	3309	1744	61%	92	±2.9%
Children	1657	593	51%	66	±4.2%
AB Adults	1376	457	60%	79	±5.6%
16-24 Adults	927	475	70%	64	±6.4%
ITV					
All Individuals	8529	2120	80%	35	±3.5%
Adults	6872	2185	81%	37	±3.4%
Housewives	3313	2227	83%	45	±3.5%
Men	3309	1744	85%	31	±4.1%
Children	1657	593	72%	23	±5.9%
AB Adults	1376	457	91%	23	±8.5%
16-24 Adults	927	475	93%	22	±8.5%
Channel 4					
All Individuals	8529	2120	89%	11	±3.9%
Adults	6872	2185	89%	12	±3.8%
Housewives	3313	2227	98%	13	±4.1%
Men	3309	1744	98%	11	±4.7%
Children	1657	593	98%	8	±8.0%
AB Adults	1376	457	90%	9	±8.4%
16-24 Adults	927	475	99%	9	±9.1%
Total Satellite					
All Individuals	8529	2120	189%	8	±8.2%
Adults	6872	2185	217%	7	±9.3%
Housewives	3313	2227	294%	6	±12.4%
Men	3309	1744	189%	8	±9.0%
					±11.2%

Children	1657	593	136%	10	±17.4%
AB Adults	1376	457	186%	4	±16.5%
16-24 Adults	927	475	180%	6	

The reductions from the actual to effective sample sizes vary according to the number of individuals per household, the degree to which the demographic group is disproportionately sampled and the range of weights required to balance other important profiles. Standard deviations tend to be more consistent between demographic groups because they are independent of sample size. However, there is some variation which is driven by the inherent volatility of the demographic group (e.g. 16-24's) and the degree of clustering within households (which smoothes standard deviations to a greater or lesser extent).

Given these variations, it is difficult to generalise with real confidence. However the standard deviations are falling into the following ranges:

	Average	**Minimum**	**Maximum**
Total TV	59%	51%	70%
ITV	84%	72%	93%
Channel 4	94%	89%	99%
Satellite	200%	136%	294%

In practice, these average standard deviations together with the actual effective samples (which are easy to calculate and are independent of the particular measurement of viewing under consideration) will give a reasonably robust working model for the calculation of sampling errors for four week average hours. It is our intention to continue the search for such practical models for the estimation of sampling errors.

5. Effective Sample Sizes

Previous studies of sampling error (e.g. Arbitron, 1974) have concentrated on comparing the sampling error for a single rating based upon a simple random sample, with the sampling error for an average rating based upon the real sample design. The relationship between the two was then used to calculate an effective sample size which incorporated all panel design and behavioural effects.

Picking up the first example from the previous section, the All Adults average hours to ITV in a single day was 1.43 hours. This is equivalent to an average minute rating of 6% (ie. 1.43 ÷ 24).

Then re-working Table 4.1 to calculate sampling errors in terms of average ratings, we get the results shown in Table 5.1.

Table 5.1 All Adults ITV Single Day			
Average Rating = 6.0%			
Complexity	**Effective Sample**	**Standard Deviation**	**Sampling Error**
Unweighted	6872	7.44	±0.090
Stratification	6872	7.44	±0.090
Weighting	4743	7.32	±0.106
Clustering	2811	6.24	±0.118
Full Specification	2185	6.36	±0.136

So for the full specification, the average rating is 6.0% with a sampling error of ± 0.136 percentage points.

Now, for a **single minute** rating of 6% from a simple random sample of 6872 adults, the sampling error can be calculated as:

$$s.e. = \sqrt{\frac{pq}{n}} = \sqrt{\frac{6x94}{6872}} \qquad = 0.286 \text{ percentage points}$$

Then if we use this same formula but insert a different value for the sample size n, say n = 30,390, we get the sampling error for a single minute rating from a simple random sample of 30,390 adults:

$$s.e. = \sqrt{\frac{pq}{n}} = \sqrt{\frac{6x94}{30,390}} \qquad = 0.136 \text{ percentage points}$$

This is the same as the full specification sampling error for a daily average rating of 6.0%. And the divisor we have used to get there, ie. n = 30,390, is another definition of the "effective sample". It is the size of the simple random sample which would give the same sampling error for a single minute rating as our actual sample gives for an average rating. Table 5.2 is an extension of Table 5.1 which shows these re-defined effective samples.

Table 5.2 All Adults ITV Single Day			
Average Rating = 6.0%			
Complexity	**Effective Sample**	**Sampling Error**	**"Effective" "Sample"**
Single Minute Formula	6872	±0.286	6872 (= x 1)
Unweighted	6872	±0.090	69395 (= x 10)
Stratification	6872	±0.090	69395 (= x 10)
Weighting	4743	±0.106	50027 (= x 7)
Clustering	2811	±0.118	40369 (= x 6)
Full Specification	2185	±0.136	30390 (= x 4)

The numbers in brackets show the factors by which the actual sample must be multiplied in order to use the $\sqrt{pq / n}$ formula to calculate sampling error. If there are simple patterns in these factors to cover a variety of audience measurements, channels and demographic groups, then we have a convenient way to calculate sampling error.

We have not pursued this alternative approach in this paper. However, it is easy to calculate this new "effective sample" for each of the audience measurements for which we have calculated sampling error.

6. Channel Share

Channel share can be based upon the viewing in a single minute or upon averages for breaks, programmes, whole days, weeks or longer. In this paper we have concentrated on shares calculated as the channel's average hours of viewing percentaged on Total TV average hours of viewing.

Channel share sampling errors are dependent upon the correlation between viewing to the channel and viewing to Total TV. For daily, weekly and four weekly average hours of viewing, we have found a positive correlation in all the examples considered. Therefore channel share sampling errors are relatively lower than average hours sampling errors.

Following the first example from section 4, Table 6.1 shows the effects of each sample design criteria on the sampling error for All-Time Channel Share to ITV for All Adults on a single day.

Table 6.1 All Adults ITV Single Day				
Channel Share = 40%				
Complexity	**Effective Sample**	**Standard Deviation**	**Confidence Interval**	**Average Hours Confidence Interval**
Unweighted	6872 *100*	83% *100*	±2.0% *100*	±3.0% *100*
Stratification	6872 *100*	82% *99*	±2.0% *99*	±3.0% *100*
Weighting	4743 *69*	81% *98*	±2.3% *117*	±3.6% *119*
Clustering	2811 *41*	73% *88*	±2.7% *137*	±3.9% *131*
Full Specification	2185 *32*	71% *86*	±3.0% *152*	±4.6% *152*

For consistency with section 4, the standard deviations and confidence intervals are shown as percentages of the channel share. The corresponding sampling errors for ITV average hours are shown for comparison and are seen to be 1½ times as big as the share sampling errors. As expected, channel share is a more robust measurement of viewing.

Table 6.2 is the equivalent analysis for All-Time Channel Share to Total Satellite for 16-24 Adults on a single day.

Table 6.2 16-24 Adults Total Satellite Single Day				
Channel Share = 8%				
Complexity	**Effective Sample**	**Standard Deviation**	**Confidence Interval**	**Average Hours Confidence Interval**
Unweighted	927 *100*	373% *100*	±25% *100*	±25% *100*
Stratification	927 *100*	296% *79*	±19% *79*	±23% *91*
Weighting	719 *78*	359% *96*	±27% *109*	±28% *109*
Clustering	591 *64*	353% *95*	±29% *119*	±30% *118*
Full Specification	475 *51*	264% *71*	±24% *99*	±29% *114*

Considering the unweighted calculation of sampling error, the confidence interval on channel share is exactly the same as for average hours. This is because there is so little correlation between satellite viewing and total TV viewing. However, stratification bites much more into the channel share sampling error. This is because the correlation between satellite and total TV only comes out when it is considered within the satellite receiving component of the panel. The effect of this correlation is still diluted by the 80% of the panel which don't

receive satellite but do contribute to the channel share Total TV base. We would expect the satellite channel share sampling error: average hours sampling error ratio to be similar to ITV if restricted to a measurement of satellite homes only.

Table 6.3 compares channel share with average hours sampling errors using a four week base.

Table 6.3 4 Week Channel Share

	Average Hours	Confidence Interval	Channel Share	Confidence Interval
ITV				
All Individuals	35	±3.5%	37	±2.1%
Adults	37	±3.4%	37	±2.1%
Housewives	45	±3.5%	39	±2.2%
Men	31	±4.1%	34	±2.5%
Children	23	±5.9%	35	±3.8%
AB Adults	23	±8.5%	30	±5.4%
16-24 Adults	22	±8.5%	35	±4.4%
Channel 4				
All Individuals	11	±3.9%	12	±2.9%
Adults	12	±3.8%	11	±3.0%
Housewives	13	±4.1%	11	±3.2%
Men	11	±4.7%	12	±3.3%
Children	8	±8.0%	12	±6.2%
AB Adults	9	±8.4%	11	±6.4%
16-24 Adults	9	±9.1%	14	±5.6%
Total Satellite				
All Individuals	8	±8.2%	8	±6.4%
Adults	7	±9.3%	7	±7.1%
Housewives	6	±12.4%	5	±9.6%
Men	8	±9.0%	8	±7.0%
Children	10	±11.2%	15	±8.4%
AB Adults	4	±17.4%	5	±13.3%
16-24 Adults	6	±16.5%	9	±12.6%

Again, channel share sampling errors are seen to be consistently lower than average hours sampling errors. The differences are seen to be greater when channel shares are higher - this is logical in that a large channel tends more to drive total TV viewing.

In absolute terms, channel share sampling errors are largest for ITV, followed by satellite, with Channel 4 being the smallest, despite the fact that Channel 4 has higher share.

7. Change Over Time

It is well known that panels are better at tracking change over time than are successive independent samples. The sampling error on change over time depends upon the correlation in viewing levels from period to period and the proportion of the total panel that is continuous. Our working formula for the sampling error on change is:

$$Sx\sqrt{2x(1-RxF)}$$

where:

S = Single period sampling error.

R = Correlation from period 1 to period 2.

F = Panel continuity from period 1 to period 2.

The panel continuity is defined to be the proportion of the period 1 sample which also reports in period 2. Continuity is decreased by the amount that an individual's weight changes from period 1 to period 2: this explains the low levels of continuity shown in the following examples.

Dependent upon the values of the correlation and continuity, the sampling error on change will be either larger or smaller that the single period sampling error. The worst possible case is when either the correlation or continuity is zero, in which case the panel measurement is then equivalent to that from two independent samples. A panel operation with enforced turnover will necessarily have lower continuity and therefore larger sampling errors on change.

Picking up on four week average hours of viewing, Table 7.1 looks at the sampling errors and their components for month on month change from March 1995 to April 1995.

Table 7.1 Month on Month Change - 4 Week Average Hours

	Continuity	Correlation	$\sqrt{2(1-R.F)}$	Average Hours	Confidence Interval
Adults					
Total TV	0.91	0.90	0.60	101	±1.4%
ITV	0.91	0.93	0.56	37	±1.9%
Channel 4	0.91	0.86	0.67	12	±2.5%
Satellite	0.91	0.95	0.52	7	±4.8%
16-24 Adults					

Total TV	0.88	0.87	0.70	64	±4.5%
ITV	0.88	0.86	0.71	22	±6.0%
Channel 4	0.88	0.84	0.72	9	±6.6%
Satellite	0.88	0.87	0.68	6	±11.2%

Panel continuity is lower for 16-24's because a proportion will leave the age group each month and a further proportion will leave the household. The serial correlations for this group are also lower than for All Adults, reflecting again their more volatile viewing behaviour. However, all correlations are high as would be expected between adjacent months.

Overall, there is a substantial gain in the precision of the panel for measuring change over time, with All Adults sampling errors being almost halved. To measure change to the same level of accuracy with successive independent samples would require sample sizes to be about six times as big.

Table 7.2 looks at the sampling errors and their components for year on year change from March 1994 to March 1995.

Table 7.2 Year on Year Change - 4 Week Average Hours

	Continuity	Correlation	$\sqrt{2(1-\text{R.F})}$	Average Hours	Confidence Interval
Adults					
Total TV	0.68	0.81	0.95	101	±2.3%
ITV	0.68	0.85	0.92	37	±3.1%
Channel 4	0.68	0.74	1.00	12	±3.8%
Satellite	0.68	0.83	0.94	7	±8.7%
16-24 Adults					
Total TV	0.50	0.61	1.18	64	±7.6%
ITV	0.50	0.67	1.16	22	±9.9%
Channel 4	0.50	0.59	1.19	9	±10.8%
Satellite	0.50	0.65	1.16	6	±19.1%

The continuity and correlations are significantly lower, resulting in sampling errors on change which are approximately the same as the single period sampling errors. So, partly due to panel dynamics (i.e. panel turnover) and partly due to behaviour (deterioration in serial correlation), estimates of short term change are more accurate then estimates of long term change: month on month sampling error is less than two thirds of the year on year sampling error.

8. Other Measurements of Viewing

In the examples explored so far, we have already seen some quite large sampling errors. In the UK market, it is also important to note that ITV is a regional broadcaster and that ITV sampling errors are about three times as large in a typical ITV region. Therefore, in terms of how the data is used, this puts ITV accuracy on a par with Satellite TV. So sampling error is a real issue which must be taken seriously by all data users. It is important to now cover some of the other key applications of the TV measurement system.

8.1 Channel Reach

Channel reach has more value to low rating channels or those with a restricted availability. One important application is to tell a broadcaster whether they need to attract new viewers or to encourage more loyalty from a large base.

We know that all-time channel reach is relatively more accurate measurement. In fact, for channels like ITV where maximum reach is approached very quickly, it is not worth showing the sampling errors. So the example in Table 8.1 is for the satellite station Sky One.

Table 8.1 Channel Reach - All Adults Sky One

	Reach %	Confidence Interval	
		Single Period	Year on Year Change
1 day	4.3	±0.6 *±15%*	±0.8 *±18%*
1 week	12.4	±0.8 *± 7%*	±0.9 *± 7%*
4 weeks	16.9	±0.7 *± 4%*	±0.7 *± 4%*

These data show that after four weeks, the reach to Sky One has built to nearly 17% - remember that Sky One is available to only 20% of the population. For this measurement, the confidence interval has reduced to a respectable ±4% for both the single four week period and its year on year change.

8.2 Programme/Commercial Break Ratings

Programme and commercial break ratings are important in their own right, but also form the building blocks which determine all-time average hours and impacts. They are also the basis on which schedules are planned. Programme and commercial break ratings carry more sampling error than all-time average hours, but in many cases users expect this and can work around it.

Table 8.2 Centre Break - Coronation Street

Monday 19:45p.m. All Adults

	Rating	Confidence Interval	
	%	Single Period	Year on Year Change
1 Monday	30.6	±1.7 *±5.5%*	±1.9 *±6.4%*
4 Mondays	30.8	±1.4 *±4.4%*	±1.4 *±4.5%*

The example shown in Table 8.2 is for the centre break in ITV's top rating programme. On Mondays, this gets a rating of just over 30% amongst All Adults.

For a single Monday, the confidence interval for the national panel is ±5.5%, which is the same order of magnitude as ITV average hours for a whole day.

The confidence interval for year on year change is somewhat higher, but not if we consider a four week average - this is probably a useful fact for a forecaster.

8.3 Schedule Reach and Frequency

Reach and frequency analysis is a key planning and evaluation tool for individual advertising schedules. Whilst the broadcaster may well have delivered plenty of impacts with the right demographic profile, it is then important to make the optimum selection of airtime for a particular advert.

The example shown in table 8.3 is a hypothetical ITV schedule in March 1995. The schedule covers a four week period and the target audience is All Adults.

Table 8.3 4 Week Schedule - All Adults

		Confidence Interval			
		Single Period		**Year on Year Change**	
ITV					
Total TVR's	484	±17	*±3.5%*	±17	*±3.6%*
1+ Cover	85.1	±1.2	*±1.4%*	±1.5	*±1.7%*
4+ Cover	51.6	±1.8	*±3.4%*	±2.0	*±3.9%*

The confidence interval for Total TVR's is pretty similar to that for four week all-time average hours of viewing. We have already seen that channel reach is a relatively accurate measurement and this example shows that the same applies to schedule 1+ cover.

The confidence interval for 4+ cover indicates the reliability of the panel in proving that a particular schedule has achieved a planned frequency. 4+ cover is seen to have about the same level of accuracy as Total TVR's.

At the national level, it seems that the panel results are fairly accurate and also make a fairly good forecasting base. Regional results would have to be treated with a degree more caution.

9. Radio Audience Research

The most widespread technique for radio audience research is the one-week personal diary. The calculation of sampling errors is subject to the same principles as outlined for television except, of course, that all the issues related to a continuous panel do not apply.

The principle of reducing variability by averaging was explored by Arbitron (1974) in terms of one-week diaries.

In the UK RAJAR has investigated the effect upon sampling error from a number of features of the UK RAJAR diary research design. The approach adopted was a small scale replication study - creating a series of matched samples from the main sample and measuring the variability which resulted. The results can be expressed in terms of design effect which is the number by which the actual sample size can be divided to generate effective sample size. A very simplified summary of results was:

A complex design involving inter-locking overlapping areas resulting in a huge number of small segments as the building bricks leads to weighting for disproportionate sampling for the segments which comprise each local stations area. This affects each area differently and although the three areas chosen for the replication study were chosen to represent the range of station types, the individual variation is considerable. National data is affected most of all in the quarters where every local station is surveyed.

The weighting effects found were for all adults:

National Station	1.6
Local Stations Range	1.0 - 1.6
Replication Study Areas	1.5

The survey design involves an element of clustering within sampling points. This is obviously particular to the UK design but the results give an idea of the kinds of effects possible for all adults:

Total Radio	Approximately	1.2
Local ILR Service	Approximately	2.1

Diaries are placed for all household members. The degree of within household correlation is less than for television but still significant, for all adults:

Total Radio	Approximately	1.4
Local ILR Service	Approximately	1.6

Combining all these factors gives total design effects for all adults:

Total Radio	2.8
Local ILR Station	4.9

In other words effective sample sizes are just over one third of actual samples for all radio and one fifth for local ILR services. Design effects for individual sub-groups are usually lower.

For radio, reach figures are subject to the sampling errors associated with a single percentage, taking into account sample design and weighting effects. For hours of listening, the appraoch discussed for television would apply. In a recent case history we calculated the sampling error on weekly average hours of viewing, based upon the sample standard deviation (assuming a simple random sample) and an effective sample adjusted by the above design effects. The level of sampling error was confirmed by calculating the month to month variability of a short time series of average hours per person, i.e. assuming that in reality audiences were stable.

10. Applying Sampling Variations to Practical Situations

10.1 A General Approach

Most uses of audience measurement are about measuring differences, changes or trends. In isolation, a single absolute audience has little practical use except in relation to other media. Mostly comparisons are made rating to rating, week to week, month to month, year to year. Superimposed on top of these are comparisons between stations and with the total audience.

In making such comparisons the danger of making reference to sampling variation or error is that it so often leads to the assumption that any difference within the 95% confidence limits is meaningless and should be ignored. This approach is fostered by academic statisticians trained in the principles of pure science who would classify only differences outside some level of confidence limits as acceptable evidence to support hypotheses. In applied science where decisions **must** be made you have to use the best evidence available. If your life depended upon it you would accept evidence with 51% confidence if nothing else were available.

Using media research, decisions about advertising or programmes have to be made and initially, it is assumed that differences between audience statistics are 'facts' and actionable. The problem is that so many such comparisons are made that some of them will not, in fact, be real. In one in twenty cases, the 95% confidence limits will be reached by chance. Given the small effective sample sizes often used in media research this can occur frequently. What then is the data user to do? We believe that there are viable positions between the extremes of blind belief in all data as facts and rejection of all interesting data changes as chance fluctuations.

We would argue that where a decision has to be made, based on the difference between two sets of audience data then an appropriate strategy involves considering the degree of risk and following different approaches for low risk and high risk decisions.

10.2 Low-Risk Decisions

An example would be choosing between two very similar advertising breaks. Using all the relevant available data, it is a sensible strategy to choose the highest rating regardless of sampling error. Generally, decisions involving the question 'which is the highest audience?' need not make reference to sampling error.

It is, however, important to use all relevant data and where low ratings on small samples are concerned then using all relevant data could involve averaging audiences over time.

10.3 High-Risk Decisions

These would tend to be more strategic decisions. Examples could be drawn from programming and scheduling; whether this month's ratings for a programme or time segment are really down on last months or, more dramatically, whether the Spring schedule this year is doing badly compared with last year. The kinds of comparisons being made in these situations are very much dependent on local broadcasting conditions. In the UK, we have a network of regional commercial stations, competing against national stations, both commercial and public service, with a variable pattern of satellite station reception. Evaluating a regional station's new Spring schedule therefore could involve regional ITV hours of viewing indexed upon ITV network hours of viewing and compared month on month or year on year. This indexing is to take account of changing national competition and the regional stations 'normal' relationship to network ratings representing regional demographics and lifestyle.

Calculation of the 'sampling error' of such statistics is quite complex, being ratios of ratios of averages. It can be done as we have shown but is it worth it? Such comparisons in the UK at least are sometimes made not just on total audience but on particular sub-groups of marketing or programming interest. Small sample sizes here mean that sampling variation can be highly relevant.

We would recommend the following strategy in evaluating the significance of such comparisons.

(i) First, consider whether there are any reasonable explanations of change in terms of programming or, more rarely, changes in the profile of the panel.

(ii) Next, consider the likely extent of sampling variation. Approximate calculations may be enough to start with. Such calculations should take account of the nature of panels. As the distance in time between the two figures being compared increases, so does the sampling error. As time passes people join and leave panels, making them a mixture of continuous panels and independent samples. Within the continuous panel people change their lifestyle and behaviour and the greater the elapsed time the lower the correlation between their behaviour at the two points in time. In other words, over time, some panel members tend to become like different people; again more like independent samples. Some of our empirical calculations suggest that sampling errors on hours of viewing year on year are nearly double those applying to month to month comparisons. A big change month to month therefore is more likely to be significant

Another factor affecting significance is than the same change year on year.whether the 'difference' under assessment, is within a sequence in the same direction representing a trend, or is a once-off step. Being part of a trend reduces the likelihood of sampling error as an explanation. Gradual trends however can be generated by gradual panel changes and an investigation of this possibility is discussed below. A difference

which is well beyond the 95% confidence limits is unlikely to be occurring by chance. Accepting this as a real change however does not exclude the possibility that a smaller real change has been exacerbated by sampling error.

Changes which are within the confidence limits, however, cannot be automatically dismissed as meaningless. Their statistical significance is a product of both the size of the change and sampling variation. If sample sizes are such that a 10% change along some dimension cannot be significant but can be important if real, then there is a real problem. Ideally, increase sample size. In practice use every bit of subsidiary information to consider whether this is a real change or not, use the best information you have and act accordingly.

10.4 Further Examination of Major Audience Changes

Sometimes audience changes are so major and not obviously explicable, that deeper investigation is called for. Within the BARB system we have found that examination of panel structure and data analyses can often throw some light on the role of sampling error.

When changes in audience levels with major strategic implications occur the following protocol of investigation would broadly apply to a commercial regional station (on which the BARB panel structure is based).

(1) Break the change down by times of day, any sub-regions and audience demographics and currently satellite vs non-satellite receivers. Establish whether the change is general, highly specific or mixed.

(2) Consider whether any programme or schedule changes or any other events can be linked to these changes.

(3) Check levels of visitor viewing and video playback.

(4) Relate the change to:

- Summary time periods to see whether it is part of a trend.
- Network data, via indexing, to see whether it reflects competition.
- Other stations to see whether the change is unique.

(5) Compare between the periods, the profiles of the panels, weighted and unweighted, attributed and against universe targets. This shows whether, to any extent, the change can be due to changing universes or panel imbalance. In parallel with this work the relationships of weighted to unweighted data can be checked to see whether there is a weighting effect.

(6) Redefine the problem in terms of times, subgroups and periods of comparison.

(7) Conduct continuous panel analyses. This involves dividing the panels into those members who reported in both periods of the comparison and the remainder representing panel leavers and panel joiners. (Definitions here can get a little tricky.)

If the changes occurs only or primarily within the discontinuous panel then this suggests that (if panel recruitment has been correctly conducted) the change is attributable to sampling error caused by reflecting a partly new panel, involving the broader type of sampling error, discussed above. Such panel changes will have been made over a period of time and could generate a gradual 'trend'.
If the change occurs in the continuous panel it is more likely that some real movement is occurring. Sampling error, however can also occur within continuous panels, via for example some behavioural or programme preference bias within the original sample selection.

The continuous/discontinuous analysis can be made for relevant subsections of the panel.

(8) From the sequence above it is possible to separate identifiable causes of audience change and quantify the residual 'unexplained' change. This residual must then be either a real change, sampling error or a mixture of both. The residual changes can then be related to confidence limits taking into account all relevant analyses above and ultimately an informed judgement made as to the relative likelihoods of real change and sampling error.

11. Choosing Sample Sizes for Audience Measurement

When setting out to design a new audience measurement system, the logical start point is to define exactly what has to be measured and with what accuracy. Then given a knowledge of the sampling errors it should be possible to determine what sample size is required. It sounds easy but in practice it isn't.

The first step is to define a comprehensive but manageable set of key audience measurements. This has to recognise the different applications such as programme scheduling or advertising sales and marketing. Sampling errors must be calculated for all key audience measurements.

A process which mustn't be under-estimated is the interpretation of the sampling error results and consideration of the accuracy really required. Apart from the fact that each group of users may well need different levels of accuracy, it is difficult to demonstrate what the potential commercial damage would be with a particular level of sampling error for a particular statistic. This is the trade-off that must be made between money at risk and the cost of the audience measurement.

Then if the cost of the preferred research vehicle gets too high, alternative methodologies will be sought. For example, a small sample meter panel may be replaced by a large sample diary survey, despite the potential measurement weakness. And at an extreme, the user may choose to go for a biased but stable measurement based upon factored or modelled data.

In the UK we are again considering the sample size requirements for audience measurement in the future. The first stage must be to get a thorough understanding of the accuracy of the current panel and how much each sample design factor affects the sampling error for each key measurement.

12. Conclusions

(1) Television panels can be regarded as measuring behaviour in a population which may differ from the total population by an unknown degree of bias.

(2) Although bias cannot be measured, sources of bias can be investigated and panel designs amended to reduce and ideally minimise bias.

(3) Within these terms, the variability of panel data can be measured in terms of sampling error about a 'true' figure for the population measured.

(4) When a television panel sample is initially selected it is subject to the full potential of sampling error just like an independent sample. This can result in a degree of panel 'bias' i.e. deviation in profile behaviour from the population generally measured by this sampling method. This 'panel bias' will endure whilst the panel remains unchanged.

The precision of an unchanged panel is much greater than that for repeated independent samples because of the correlation between members' behaviour over time.

As panel membership changes the recuitment of new members adds to sampling error, but potentially varying any 'panel bias'. Some continuous members of the panel change lifestyle and viewing behaviour over time. The decreasing serial correlations in the behaviour of continuous members behaviour and the replacement of some panel members generates increasing sampling errors for comparisons over time. If a policy of enforced panel turnover were adopted the sampling error for comparisons between two points in time would be as for two independent samples.

(5) The various elements of panel design such as stratification, sample clustering, clustering within the home and weighting need to be taken into account when calculating sampling errors.

(6) The use of television audience measurement data involves masses of comparisons between different figures. Inevitably, for 5% of these comparisons, differences will arise by chance alone which fall outside the 95% confidence limits. It is therefore important to have some idea of what can occur by chance.

(7) Approaches to calculating sampling error for the different kinds of figures used in advertising trading and programming have been described:

Single Ratings
Averages of Ratings
Hours of Viewing

Audience Share
Channel Reach
Schedule Analyses

The added complexity of comparisons over time has also been discussed.

(8) Sampling error cannot be used to suggest that all differences that lie within it are meaningless. Indeed many real, actionable differences probably will lie within 95% confidence limits.

Therefore a general philosophy has been advanced on how to use sampling error.

In extreme cases detailed analyses of panel data can help to understand the role of sampling error.

(9) In general, data users should be aware of sampling errors but still work on the data as the best information available. Where surprising results occur not supported by the context of data or broadcasting conditions a more detailed appraisal of the role of sampling error is recommended.

Further Investigations

It is proposed to calculate an extensive range of sampling errors for different data comparisons within the BARB system. It is then hoped, via further simplifying assumptions, to evolve some simple reference tables with appropriate guidance as to their use.

References:

American Research Bureau Inc. New York, (1974),
Arbitron Replication: A Study of the Reliability of Broadcast Ratings.

Keith Boyd, Simon Godfrey, Tony Twyman Market Research Society Annual Conference, Brighton, UK, (1979), The Accuracy of Hall Testing in Relation to Marketing Decisions.

Results of a JICTAR Study' in BARB Television Reference Manual, London, (May 1993),
Broadcasters Audience Research Board: 'Averaging Ratings Reduces Sampling Error:

Research Services Limited: 'Variations in RAJAR Audience Estimates' (unpublished) London, June 1995.

The Authors:

Tony Twyman is Technical Advisor to BARB: Broadcaster's Audience Research Board and AIRC: Association of Independent Radio Companies, UK

Steve Wilcox is Technical Director at RSMB Television Research Limited, UK