BROADCAST AUDIENCE RESEARCH COUNCIL INDIA

What India Watches™

Data Processing and Validation Processes

Rev1

March 2022

Table of Contents

Executive Summary

The quality of audience estimates is assessed through accuracy and precision. These are two important, yet distinct factors. Accuracy has to do with the quality of measurement coming directly from the meters and precision has to do with the degree of variability arising from the sample relative to the population of interest. BARC India takes both constructs quite seriously and has built a complex data retrieval and validation process to ensure that BARC India's audience estimates reflect "What India Watches™".

BARC India's processes are built upon two key principals: data security and empirically supported statistical methods.

This paper reviews in detail the various steps which the 10 petabytes (PB) of data, processed annually, flows through from the meters themselves to the final estimates delivered to clients. All processes are systematic and flow through a stepwise process of automated and SOP driven procedures. Several checks and balances are in place to ensure that the integrity of the data, and subsequent audience estimates, remain intact and statistically sound.

The first section of the paper focuses on BARC India's data retrieval and security processes and details how the data is collected and transferred from the panel households to BARC's data validation processes. These steps involve collection of viewing records through audio watermarks by the proprietary BAR-O-Meters. The data flows through several secure internal systems and is processed through multi-layered security systems. This data possesses all of the defining characteristics of Big Data: Volume, Velocity, and Variety.

Since the data is derived from a sample, albeit a quite large sample, it can be subject to the impact of the statistical concept known as "Outliers". This paper reviews what is an Outlier and summarises BARC's processes that not only identify, but also how outliers are treated. The Data Validation procedures follow a four-step sequential process: (a) the identification of forced viewing due to landing pages; (b) quality control procedures; (c) respondent level outlier detection and treatment; and (d) channel level outlier detection and treatment.

While these processes are complex, an effort has been made to explain them in detail in a clear and transparent manner. It is hoped that this will aid in a better understanding of BARC India's production of audience estimates and the rigour and empiricism that is systemically invested in the process.

Data processing at Broadcast Audience Research Council (BARC India)

Technology: The Backbone of BARC

With Measurement Science at its heart, futuristic technology acts as the backbone for BARC India. This technology ensures that BARC India collects desired viewership and audience data from the recruited panel homes to support its analytical needs through more than 45,000 Internet-of-things (IOT) devices known as BAR-O-Meters.

Every Technology adoption, implementation and major decision goes through Technical Committee (TechComm) for review, and recommendations are incorporated.

For more information on data collection process refer to https://barcindia.co.in/technology.

System Overview

The BARC system consists of several separate steps (Figure 1). Technology is the key enabler for Panel Management, Big Data Management, and Customer Relationship Management for both Meterology Data Limited (MDPL) and BARC's Measurement Science department. These are essential parts of panel design, recruitment, and installation/maintenance. The key components are described in further detail:

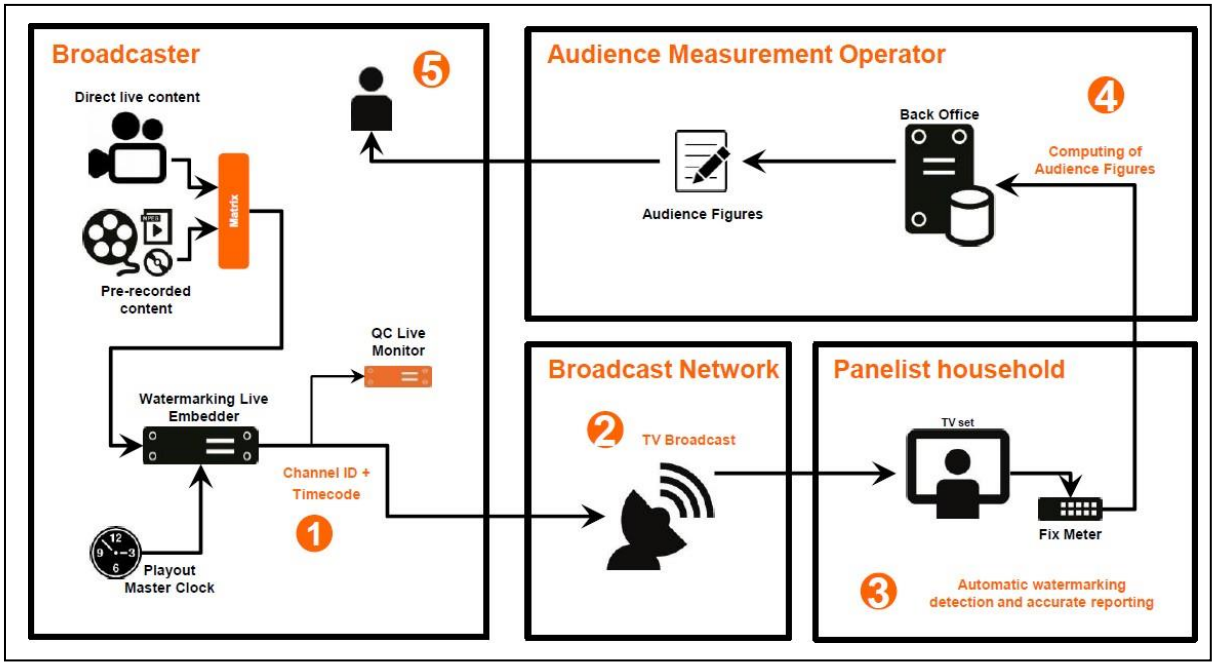


Figure 1. BARC – System Overview.

Audio Watermark

Audio Watermarking (WM) embeds unique identifiers along with timestamp in the audio of each channel prior to upload and broadcast. These encrypted watermarks are not audible to the human ears but can be detected and decrypted using dedicated hardware and software algorithms. The Audio Watermarking technology is adapted and implemented as per the TechComm recommendations.

KANTAR is the WM technology partner for BARC, and Broadcasters are responsible to implement WM in their feed itself to ensure BAR-O-Meters detect them efficiently.

Playout

The audio watermark enabled content is broadcast by the channel via satellite and on ground the signal is redistributed via a Direct-to-Home Operator (DTH) or Cable Multi System Operator (MSO).

BAR-O-Meters

The IOT devices (i.e., meters installed at recruited panel homes basis recruitment) can detect the audio watermark present in the channel while it is being viewed at a household (HH) level. The meter detects HH level data (i.e., what is being viewed at what time) by decrypting the Watermark identifier and the timestamp.

Along with the meter, a remote-control unit is provided to each HH for individual button pressing which allows an understanding of the audience profile (i.e., individuals) during viewership. This unit is securely coupled with the respective BAR-O-Meter. This step enables BARC to know who is watching the TV.

The BAR-O-Meters are indigenously designed, developed, and manufactured in India with various partners. Events from these meters are sent to BARC's back-office, an Amazon Web Services (AWS) cloud-based application, enabling the collection of events from the meter at regular intervals via data communication (i.e., 2G, 3G or 4G connections).

Viewership Events

The BAR-O-Meters are enabled with a return path (i.e., SIM card enabled) for sending the viewership events to the back-office. Both the meters and the back office are registered Ips of BARC.

Panel to Product Data Process Overview

BARC's data goes through several phases through its technological journey (Figure 2). Annually, BARC processes 10 peta-bytes (PB) of data (i.e., 1 PB = 10^{15} bytes) which is the equivalent amount of data as 667 times that of the US Library of Congress (estimated by the US Library of Congress at 15 TB) or 1,10,000 4K movies. BARC's data is truly the definition of "big data" and has all the components of the three Vs of big data: (a) Volume; (b) Variety; and (c) Velocity. The key stages in the flow are described in further detail.

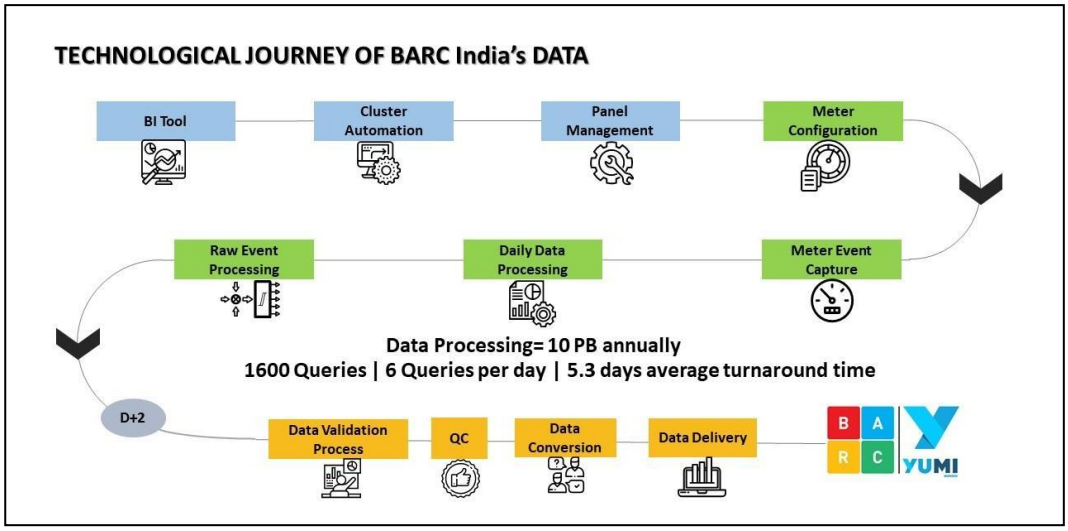


Figure 2 The Technological Journey of BARC's Data.

Panel Management

Technology enables automated, secure onboarding and lifecycle management. The applications are integrated with internal systems for intelligent task scheduling and tracking as per the Standard Operating Procedure (SOP) and Turn Around Time (TAT) defined by MDPL.

Big Data Management

The events generated (i.e., WHO is watching WHICH channel at WHAT time), by the BAR-O-Meter at a HH level and the individual button pressing, are collected, and processed at the back-office level. Viewership data is cleaned, merged with channel, program, language, and broadcast schedule details. Universe estimates are applied to get the projected viewership data. Technology runs this process daily, based on the automated jobs scheduled and the logics and algorithms defined by Measurement Science and approved by the BARC Technical Committee (TechComm).

Further details can be found in BARC's Description of Methodology on the website: https://barcindia.co.in/measurement/television-audience-measurement-description-of-methodology.pdf

Customer Relationship Management

The currency data (i.e., viewership data) processed, is made available to the registered end users via secured lines on a daily and weekly basis as per the product (e.g., SPOTTREK, YUMI). The customer-centric data analytics tool is used by the broadcasters, advertisers, and agencies to do the further planning and data-driven decision making.

Multi-layered Security Model

To ensure that the BARC panel, meter events, and currency data is safe and securely processed and stored, BARC has enabled a best-in-class security framework at our data centre (DC) and cloud applications (Figure 3). BARC performs 3rd party audits around vulnerability assessment and penetration testing (VAPT), processes via Centre d'Etude des Supports de Publicite (CESP), and technical audits to take feedback and strengthen the ecosystem basis those. A lambda (λ) architecture (i.e., DC plus cloud) has been deployed ensuring 99.999% uptime and scalability for this complex industry solution. The key components are described in detail below:

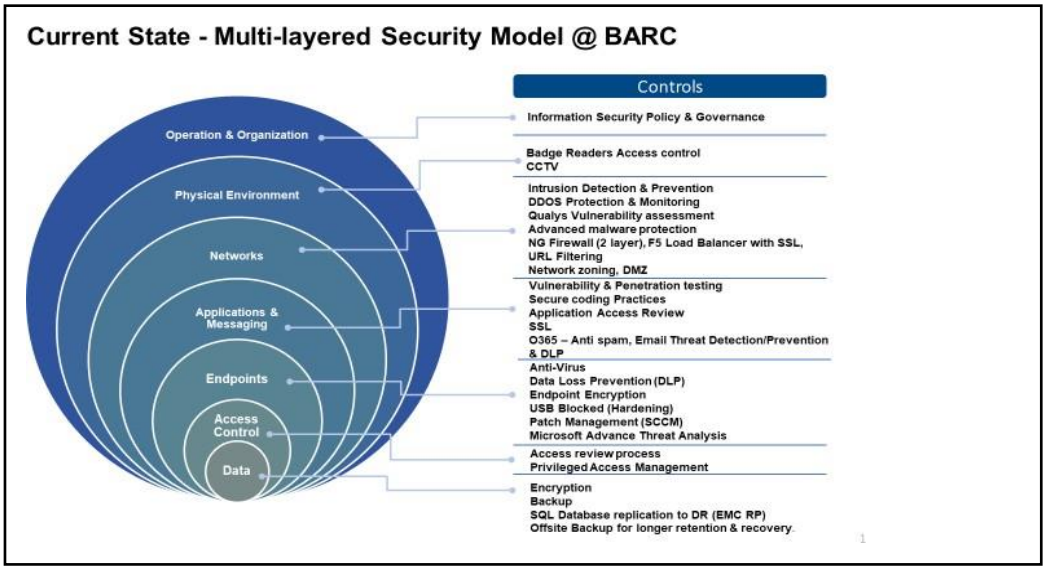


Figure 3 Multi-layered Security Model

Infrastructure

BARC adapts hybrid infrastructure with physical data centre and cloud services across its application and software (Figure 4).

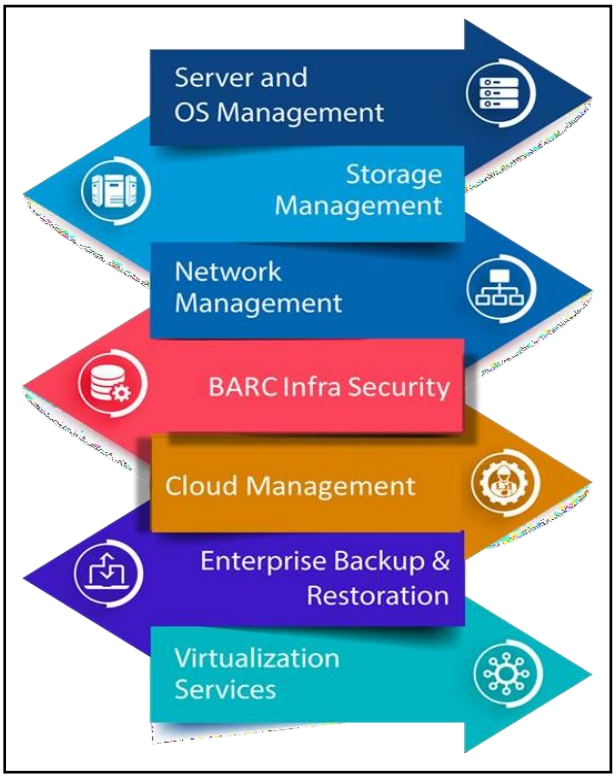


Figure 4. BARC's Hybrid Infrastructure

Security (Data Ce & Cloud)

The key components of BARC's data security are as follows:

- A fully secured Data Centre.
- Three tier security implemented. Perimeter firewall plus internal firewall & Physically isolated DMZ cluster.
- IPS, IDS & Antivirus scanning for all the packets on firewall.
- All the URLs bind with SSL certificate with SHA256 encryption.
- Privilege Identity management implemented for all the server and database access, which logs each and every activity.
- DDOS services activated with 24x7 monitoring.
- AWS cloud for different workload coming with AWS inbuilt security.
- Standard and best practice of AWS to Secured data on AWS S3 with IAM & access key control access.
- Accessed through "Privilege identification Manager" (email + SMS OTP based) PIM & Jump server for AWS; and
- "TAC Securities" completed the VAPT audit for applications.

Communication Security

The BARC meters distributed in the field, send data through TLS (Transport Layer Security) over TCP/IP to the workers-modules hosted on cloud. TLS is a suite of protocols to provide secure communication as follows:

- Confidentiality is maintained by applying block & stream ciphers
- Integrity with Message Authentication Code (MAC)
- Authenticity with certificates

The TLS protocol provides secure communication on the Web. All data is encrypted using a 256-bit encryption. 256-bit encryption is a data/file encryption technique that uses a 256-bit key to encrypt and decrypt data or files. It is one of the most secure encryption methods and is used in most modern encryption algorithms, protocols and technologies including Advanced Encryption Standard (AES) and SSL (Secure Sockets Layer).

256-bit encryption refers to the length of the encryption key used to encrypt a data stream or file. A hacker or cracker will require 2^{256} different combinations to break a 256-bit encrypted message, which is virtually impossible to be broken by even the fastest computers. It would take far longer than any of our lifetimes to crack an AES 256-bit encryption key using modern computing technology.

The TLS protocol specifies a well-defined handshake sequence to perform this exchange of data. As part of the TLS handshake, the protocol also allows both peers to authenticate their identity. It uses public key cryptography (also known as asymmetric key cryptography), which allows the peers to negotiate a shared secret key without having to establish any prior knowledge of each other, and to do so over an unencrypted channel. SSL/TLS is the standard technology for keeping an internet connection secure and safeguarding any sensitive data that is being sent between two systems, preventing criminals from reading and modifying any information transferred, including potential personal details.

Finally, with encryption and authentication in place, the TLS protocol also provides its own message framing mechanism and signs each message with a Message Authentication Code (MAC). The MAC algorithm is a one-way cryptographic hash function (effectively a checksum), the keys to which are negotiated by both connection peers. Whenever a TLS record is sent, a MAC value is generated and appended for that message, and the receiver is then able to compute and verify the sent MAC value to ensure message integrity and authenticity.

Data validation at BARC

Data validation is an important part of estimation of population parameters (e.g., the TV viewing of Indian individuals) through samples such as the BARC currency TV panel or even census, or census-like, data such as the Indian Census or return-path data (RPD) systems. Data validation is supported through the statistical literature and is conducted by many organisations involved in research of many forms – including the work done by the Data Processing Division of the Office of the Registrar General & Census Commissioner, India, for the Indian Census.

This section outlines what are statistical outliers and why they are an important consideration for data validation followed by an overview of the data validation processes employed by BARC.

Statistical Outliers

An outlier is a statistical construct with respect to a dataset. Simply speaking, an outlier can be described as an observation that lies an abnormal distance from other values in a random sample from the population, such as BARC's viewership data captured from its currency TV electronic measurement panel. An outlier can manifest at various levels or aggregates of data such as individual level data, household level data, or even channel level data.

An outlier may take two forms:

1. A representative outlier: An outlier which cannot be regarded as unique in the population. That is to say that there are likely other individuals, households, or channels in the population which behave in the same manner. These are considered valid observations and should be kept in the sample.
2. A non-representative outlier: An outlier which can be regarded as unique in the population. That is to say that there is likely no other, or extremely few (i.e., well below the projected weight), individuals, households, or channels in the population which behave in the same manner. These observations will impact estimates and result in a bias.

While there is no single agreed upon approach for either the identification or treatment of outliers, there are many generally accepted statistical methods. Identification methods are well researched and constantly evolving. Some of the more common methods would include developing thresholds using interquartile ranges (IQR), Median Absolute Deviation (MAD), or the sample mean and variance. The treatment of outliers can typically take one of five approaches (Table 1)

Table 1

Approaches for the Treatment of Outliers	
Approach	Description
Retention	The outlier is kept as-is in the data set and not altered or removed
Winsorizing	*Winsorizing can take two forms: capping or substitution*
Capping	The outlier is capped at a pre-determined level based on the assumed statistical distribution
Substitution	The outlier is replaced with the nearest 'non-suspect' observation in the dataset
Exclusion	Also known as trimming, the outlier is discarded from the dataset
Reconsider	Reassessing the assumed data distribution and reassessing if the outlier continues to be unusual under a new assumed distribution

What Is Data Validation and Why Is It Important?

Audience estimates are a product of the underlying data captured by BARC's currency TV electronic measurement panel. Accuracy and precision are often used as the means in which the quality of an estimate derived from a sample is measured. These are sometimes also understood, or referred to as, validity and reliability. Both constructs refer to types of errors associated with the estimate of interest. Accuracy focuses on systematic errors in measurement – or biases. These could be biases

due to incomplete sample frames (e.g., the exclusion of households in rural India), biases due to technological limitations (e.g., an audio stream is required to capture an audio watermark), or processing errors. Precision focuses on the error from only observing a part (i.e., sample) of the population – often referred to as sampling error – where the sample does not perfectly represent the population. In certain cases, precision can be measured through the standard error.

These can be easily understood using the analogy of a dart board (Figure 5). Accuracy refers to how close the darts fall to the bullseye (i.e., the target), whereas precision refers to how consistently close the darts fall to one another. A darts player can be either accurate or precise, both accurate and precise, or neither accurate nor precise.

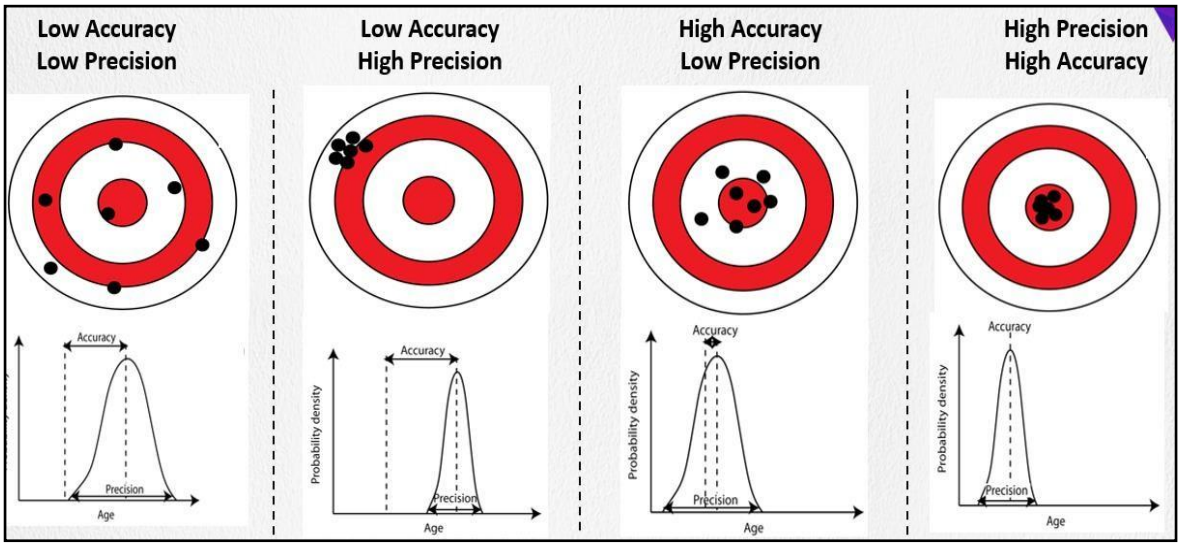
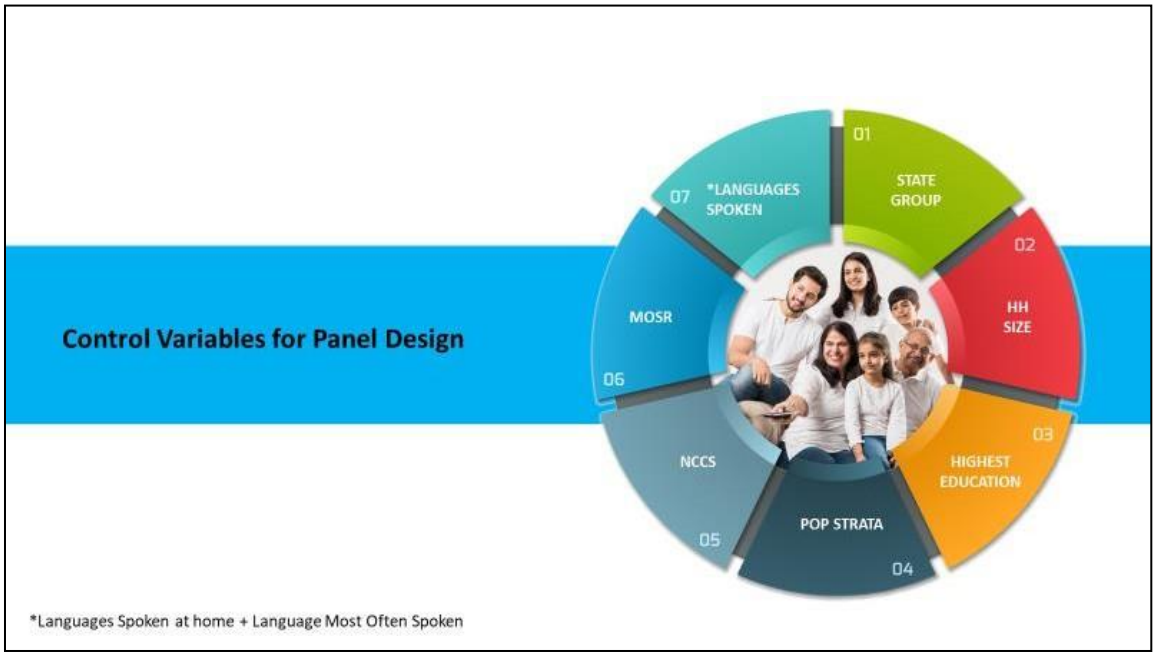


Figure 5. What's the Difference Between Precision and Accuracy?

Outside of technology and data production issues, accuracy is typically controlled through a strong sample design and sampling plan. Precision, on the other hand, is typically controlled through sample size where larger sample sizes, all other things equal, tend to produce more precise estimates than smaller samples; however, once large samples sizes such as BARC's panel of 44,000 households are achieved, the gain to precision from additional sample becomes small.

The panel itself is matched to the population to ensure it is representative with respect to key variables which have been shown to be the most effective in capturing the variation in television viewing (Ref. Image below). If the panel is balanced to these variables, the distribution (i.e., viewing profile) of the underlying panel should match the distribution of the population. However, if certain abnormalities exist in the underlying panel data, this expected result will not hold true.



Control Variables

This can be demonstrated through a simple example. Assume that a survey has been set to estimate the average height of males. Ten males are randomly selected and asked to submit their height as measured through a measuring tape. The results are collated and from the sample a diagnostic check (MAD) is conducted to identify potential outliers (Figure 6). It is seen that the resulting data obtained from male number 3 could be an outlier at 65.7 cm. Including this data point in the calculation would result in the estimated height being 167.3 cm as opposed to 178.6 cm without. This is a difference of -6.3% in the estimated height.

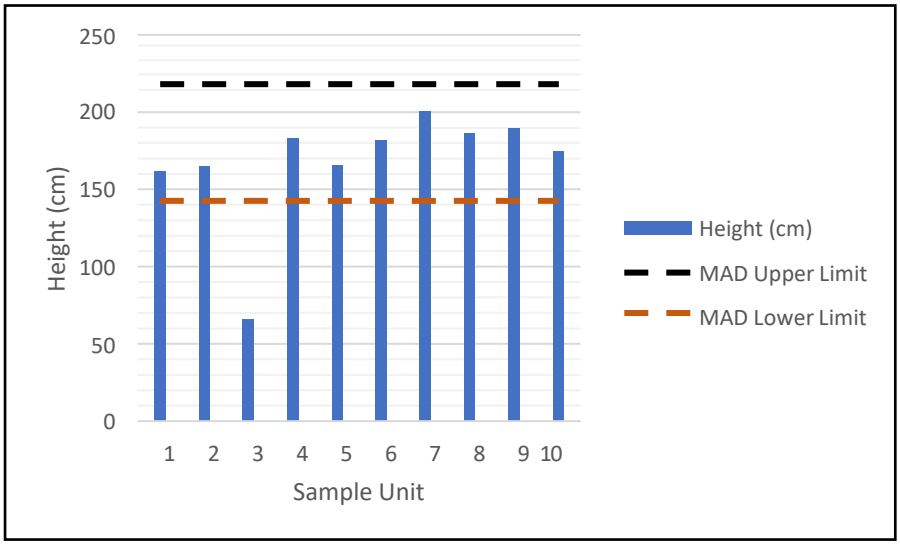


Figure 6. Example Survey Data

Further investigation on Male number 3 showed that in error, he had submitted his height in inches rather than centimeters. Since the issue was identified, new data could be collected, and a new average could be estimated (Table 2). It should be noted that the estimate with the outlier excluded, as well as the estimate based on the revised data, are very similar.

Table 2

Comparison of Estimated Average Heights	
Estimate Procedure	Estimated Average Height
Original data with outlier included	167.3
Original data without outlier included	178.6
Revised data	177.4

It is therefore easy to understand based on the above analogy how the inclusion of outliers in the data used to estimate television audiences could result in errors, or more specifically, bias in the audience estimates. It is therefore imperative that BARC review the data, and treat outliers as required.

Potential Causes of Outliers in BARC Data

There are a variety of reasons why outliers may manifest in BARC's TV audience estimates (Table 3).

These reasons could have significant impact on the accuracy and precision of BARC's audience estimates if ignored. Therefore, BARC's data validation process has been implemented to identify the various outliers, treat as statistically appropriate, and flag for technical, vigilance or panel intervention as required – reducing the likelihood of the same phenomenon from the individual, household, or channel being repeated.

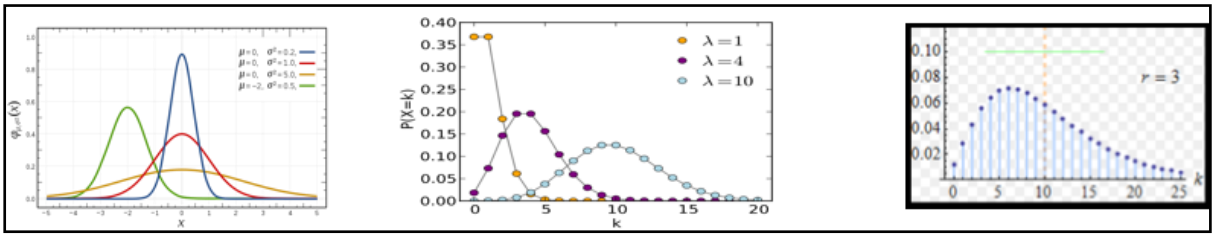
Table 3

Examples of Causes of Outliers in BARC Data		
Cause	Type	Considerations
Extreme Viewer	Representative	• Valid observations but increase variability • Their removal would cause bias • Treatment becomes a trade-off between variability and bias • Difficult to assess if these are representative cases or due to other reasons below
Poor Household Compliance – Intentional	Non-representative	• Typically manifests as unusually long periods of continuous viewing or strange patterns of viewing within the household • Thresholds help, but other mechanisms need to work in parallel • Unintentional cases can often be coached and allowed to remain in the sample
Poor Household Compliance – Unintentional	Non-representative	
Meter Capture Issues	Non-representative	• These can manifest as channel being understated in audience estimates, or absolutely no viewing being captured
Data Transmission Issues	Non-representative	• Meter is operating, but data is not being transmitted to BARC resulting in higher non-connectivity and thus impacting audience estimates
Panel Infiltration	Non-representative	• Difficult to identify all cases and is often not isolated to a single household • Can manifest in high ATS • Household level manifestation rather than individual level • Impacts other channels by misdirecting viewing of the panel.

Considerations with Respect to Outliers in TV Data

There are a few important considerations with respect to outliers in TV data.

1. While outliers and their identification and treatment are well documented in the statistical literature, there is no single consensus.
2. Many common approaches to outlier identification assume a normal distribution. Television viewing is typically right skewed following more of a pattern similar to a Poisson or Negative Binomial distribution (Figure 7).
3. Therefore, approaches must either be non-parametric or use average values of viewing which will follow a normal distribution;



Ref. Image 1 *Ref. Image 2* *Ref. Image 3*

Figure 7. Comparison of a Normal Distribution, A Poisson Distribution and a Negative Binomial Distribution.

4. Theoretically outliers can be right-tailed (i.e., high values) or left-tailed (i.e., low values). Since television viewing is bounded by zero (i.e., cannot have negative viewing), TV outlier approaches tend to focus only on high viewing cases.
5. Outlier behaviour can be linked to compliance issues and therefore data quality control is required; and
6. Panel infiltration is a reality and can manifest in several ways:
    a. Artificially inflating a channel's audiences;
    b. Artificially decreases the viewing of other stations, drawing viewing away from them, in the infiltrated household.

    Therefore, infiltrated household must be dealt with through substitution or exclusion. BARC removes all households identified as being infiltrated from the panel.

BARC's Data Validation Process

There are four primary steps to BARC's data validation process: (a) Landing Page Algorithm (LPA); (b) Phase I – Data QC; (c) Phase II – Respondent Level; and (d) Phase III – Channel Level. These are described below in details.

(a) Landing Page Algorithm (LPA)

At the request of the industry, BARC launched the LPA in September 2020. The LPA identifies viewing that occurs as a result of a channel being placed on a landing page. This identification is based upon seven different statistical checks, comparing the viewing of a channel when it is the first viewed in a viewing session to when the viewing to the channel occurs in latter parts of the viewing session. The LPA has been tuned in order to maximize the success rate while minimizing false positives (i.e., where a channel is incorrectly identified as being on a landing page).

Once the viewing has been identified as being due to a landing page, the LPA then removes only the portion of viewing which is deemed to be forced and where the viewer is likely not engaged. The

balance of viewing is kept. This, again, is done by comparing viewing markers of the channel when it is served via a landing page to when it is viewed naturally.

The LPA is fully automated and is system driven. There is no human intervention.

(b) Phase I – Data QC

BARC ensures frequent coaching and training of panel households with respect to the meter as well as the button pushing. Button pushing is what allows BARC to transcribe household level viewing to individual level viewing and, therefore, it is critical to have a high level of household compliance. Despite ongoing training, there are occasions where the household may not properly comply. These cases result in extremely long viewing sessions, overnight viewing sessions and many other situations. It is, therefore, necessary for BARC to identify these sessions and adjust accordingly as a part of data Quality Control (QC).

Four scenarios are assessed against empirically supported thresholds. Viewing that exceeds these thresholds is removed from the in tabulated base prior to weighting – ensuring that the weights of the intab individuals are properly projected to the population. These thresholds are empirically reviewed annually and adjusted should the viewing patterns change.

The Phase I process is fully automated and is system driven. There is no human intervention. The changing of threshold values requires the approval of the BARC Oversight Committee (OC).

(c) Phase II – Respondent Level

Respondent, or individual, level outliers are detected and treated by using a statistical metric known as Median Absolute Deviation (MAD) and comparing each individual's viewing behaviour against: (a) their past viewing behaviour; (b) the viewing behaviour of other like individuals; and (c) control groups behaviour of the day with past viewing behavior. Data is analyzed at the Channel level as well as genre level.

Individual viewing that is identified as being an outlier is adjusted through a statistical process known as *Winsorizing* where the viewing is kept but capped to the MAD threshold.

The Phase II process is fully automated and is system driven. There is no human intervention. The changing of threshold values requires the approval of the BARC Oversight Committee.

(d) Phase III – Channel Level

Channel level outliers are detected and treated by using a statistical metric with respect to the channel's Average Time Spent. Empirically supported thresholds and analysis of the mean and standard deviation of the channel's performance, with respect to the performance of other channels within the genre, is conducted.

Channels that are identified as being an outlier are adjusted through a statistical process known as *Winsorizing* where the viewing is kept but capped to thresholds that are linked to the patterns of similar channels within the genre.

Since there are often many reasons why a channel, or several channels, may spike in viewership due to content related events (e.g., World Movie Premiere (WMP) for Movie Genres, Heavy News Days for News Genres), it is imperative that there is a mechanism to override Phase III. *Note: While Phase III can be over-ridden, the over-ridding requires Oversight Committee (OC) approval. At the same time,*

the reverse (i.e., applying Phase III when the Channel has not been identified as an outlier) cannot occur and cannot be requested by either BARC management or the OC.

The Phase III is driven through a fully automated algorithm. All thresholds, treatments, as well as exception decision criteria, are documented and processed through a fully automated system, and the analytics team reviews the data at multiple levels with two layers of checking, before being presented to the OC weekly prior to the weekly data release. In order for Phase III to be overridden for a genre and/or channel, supporting content/event reasons along with supporting data, must be presented to the OC for their approval. For over 99% of the cases, the override occurs at a genre level for a particular day and region.

Bibliography

Figure 5: What's the difference between precision and accuracy?
Bethan Davis [22.6.2020]. **Precision and accuracy in glacial geology**
Source: http://www.antarcticglaciers.org/glacial-geology/dating-glacial-sediments-2/precision-and-accuracy-glacial-geology/

Figure 7: Comparison of a normal distribution, a Poisson distribution and a negative binomial distribution

Normal distribution (Ref. Image 1)
Public domain [22.04.2008]. **A selection of Normal Distribution Probability Density Functions (PDFs). Both the mean, μ, and variance, σ^2, are varied. The key is given on the graph.**
Source: https://en.wikipedia.org/wiki/Normal_distribution; https://en.wikipedia.org/wiki/Normal_distribution#/media/File:Normal_Distribution_PDF.svg

Poisson distribution (Ref. Image 2)
Skbkekas [10.02.2010]. **Plot of the probability mass function for the Poisson distribution.**
CC-BY-3.0. Self-published work
Source: https://commons.wikimedia.org/w/index.php?curid=9447142

Negative binomial distribution (Ref. Image 3)
Public domain [18.01.2010]. Probability mass function of a negative binomial distribution
Source: https://en.wikipedia.org/wiki/Negative_binomial_distribution

Winsorizing or **Winsorization** is the transformation of statistics by limiting extreme values in the statistical data to reduce the effect of possibly spurious outliers. It is named after the engineer-turned-biostatistician Charles P. Winsor (1895–1951).