# Exploring Nonprobability Methods with Simulations from a Common Data Source: Culture and Community in a Time of Crisis

Jennifer Benoit-Bryan[1] and Edward Mulrow[2]
[1]Slover Linett Audience Research, 4147 N Ravenswood Ave, Chicago, IL 60613
[2]NORC at the University of Chicago, 4350 East-West Highway, Bethesda, MD 20814

**Abstract**
Survey researchers have proposed approaches to estimation from nonprobability samples and have examined properties of estimators using case studies and Monte Carlo simulations. For this JSM session, we have asked three groups of researchers to apply their methods on a common set of simulations. This paper provides background on the data source, Culture and Community in a Time of Crisis (CCTC), and the design of the Monte Carlo simulation. CCTC is one of the largest surveys of cultural attendees ever conducted with over 100K respondents. It contains a wealth of behavioral and attitudinal observations that makes it ideal for simulation purposes. We go over the steps used to construct probability and nonprobability samples from CCTC records that act as the population from which both types of samples are selected. Additionally, we look at the bias inherent in the simulation design for the collection of variables researchers used to evaluate nonprobability methods.

**Key Words:** Nonprobability sample, Monte Carlo simulation

## 1. Introduction

Survey researchers have proposed three general modeling approaches to estimation from nonprobability samples: quasi-randomization, superpopulation modeling, and doubly robust. Through case studies and Monte Carlo simulations, researchers have used different data sources and response variables in their evaluations of various estimation methods, making the properties of the estimators less comparable than they can be. The distinctive feature of the JSM 2021 session "Interpreting Nonprobability Samples: Discoveries and Challenges" is that research teams presenting at the session agreed to use the same simulation data, and focus on the same set of response variables, to evaluate methods for utilizing nonprobability samples.

This paper provides background on the data source used for the simulation, and the details of how the simulation was constructed. Section 2 of this paper describes Culture and Community in a Time of Crisis (CCTC), a special edition of Culture Track, which is a survey of the public and culturally active Americans. Section 3 provides background on the simulation design for probability and nonprobability samples. Section 4 is an evaluation of the known bias one would have if the nonprobability samples were simple random samples, and Section 5 provides some concluding remarks on the usefulness of the simulation for evaluating estimation methods.

## 2. Data Source: Culture and Community in a Time of Crisis

The core goal of the CCTC research was to keep the cultural sector in dialogue with its communities and participants during the pandemic, with a focus on informing deeper equity and justice throughout the sector. The study—one of the largest surveys focused on arts and culture ever conducted to date—explored a range of challenging questions about the role and relevance of arts and culture in the lives of Americans as well as the kinds of changes people want to see in these organizations in the future. CCTC Wave 1 was fielded between April-May of 2020. CCTC surveyed more than 120,000 Americans online with the help of 653 organizations across the arts and culture sector who all distributed the survey to a portion of their lists. The survey was also distributed via the NORC AmeriSpeak panel, collecting an additional 2,027 responses.

"Key Findings from Wave 1" was published in July 2020 and has been accessed by more than 17,000 users from around the globe. Separately, the first wave of survey data was analyzed by the Slover Linett Audience Research team through the lens of race and ethnicity (Benoit-Bryan, 2020), the result of which is a policy report focused on Black, Indigenous and people of color—or BIPOC Americans: "Centering the Picture: The Role of Race & Ethnicity in Cultural Engagement in the U.S." Microsoft has enabled ongoing access to an Interactive Tool that allows anyone to explore the Wave 1 survey responses. The tool enables comparisons by genre of cultural institution, as well as race and ethnicity to further aid cultural organizations in planning and implementation toward increased relevance, equity, and sustainability.

Because of its large size, both in terms of the number of response (123,757) and the number of variables in the public use data set (693), and collection of both behavioral and attitudinal variables, we consider CCTC an ideal data source for simulation purposes.

## 3. Monte Carlo Simulation Design

The basic idea of the simulation is to select samples from a well-defined population of records using different sample designs. Estimation methods can be evaluated by comparing characteristics of the estimates with known population values.

The simulation consists of **1,000 iterations**. For each iteration, a **probability sample of size 1,000** was selected, and a **nonprobability sample of size 4,000** was selected. The details of the samples we designed are provided below.

### 3.1 Frame Construction

Since our interest is to evaluate estimation methods that utilize nonprobability samples, we need to develop a way to mimic sample selection problems that one might encounter when utilizing nonprobability samples in practice. We have chosen to mimic the type of coverage bias typically exhibited in online opt-in nonprobability samples. Such panels are known to not provide broad coverage of the US general population. For one, the online nature of the panel means that only those with internet access are in the panel. Additionally, since these are opt-in panels, there is little recruitment done to try to make the panel representative of all US internet users. Nonetheless, such panels are prevalent, and they provide low-cost alternatives to well-designed surveys of the population. So, the nonprobability samples in the simulation are selected from a subset of the full population. Similar approaches were used in Yang et al, 2019, and Valliant, 2020.

To do this, we have two sampling frames, one a subset of the other. Frame 1 is the full population of records. We use CCTC Wave 1 since it has a large set of records. We modified CCTC Wave 1 in two ways: 1) we removed records for non-US respondents, and 2) we removed records for which the census division code was missing. This reduced Frame 1 to 113,549 records, but it allows us to select stratified samples using common US geographic features.

For Frame 2, which will be the frame from which nonprobability samples are drawn, we removed records from Frame 1 in two ways. First, we removed a set of records based on descriptive variables in the file. The remaining set of records was then sorted based on other variables in the file, and additional records were removed by selecting row numbers using a highly skewed binomial distribution for which low row numbers have a high probability of selection. This two-stage removal process results in a nonprobability frame with 74,202 records. Research teams were not told which characteristics were used to remove records because this type of information is not known in practice.

Most CCTC outcome variables are categorical, and the estimates of interest are proportions. For each outcome variable of interest, let $P_{Frame\ 1}$ and $P_{Frame\ 2}$ be the population proportion computed from the probability and nonprobability frames, respectively. The **known absolute bias** associated with the nonprobability frame (Frame 2) under-coverage, $B_{pop}$, is calculated as the difference of population proportions between Frames 1 and 2.

$$B_{pop} = |P_{Frame\ 1} - P_{Frame\ 2}|$$

### 3.2 Sample Selection

Probability samples were selected from Frame 1 using a stratified design. Eighteen strata were defined by census division and Metro/Non-metro status. Samples of size 1,000 were selected using proportional allocation with a minimum stratum sample size of 40. Table 1 provides the population counts and probability sample size within each stratum.

**Table 1**: Frame 1 counts and probability sample sizes for each census division and metro/non-metro status.

| | METRO | | | |
|---|---|---|---|---|
| | Frame | | Probability Sample | |
| Division | Non-metro (0) | Metro (1) | Non-metro (0) | Metro (1) |
| 1 | 1,243 | 7,148 | 40 | 41 |
| 2 | 665 | 19,159 | 40 | 111 |
| 3 | 1,111 | 15,748 | 40 | 91 |
| 4 | 796 | 7,439 | 40 | 43 |
| 5 | 1,498 | 21,062 | 40 | 121 |
| 6 | 428 | 2,571 | 40 | 40 |
| 7 | 498 | 8,682 | 40 | 50 |
| 8 | 656 | 6,444 | 40 | 40 |
| 9 | 724 | 17,677 | 40 | 103 |

Nonprobability samples were selected from Frame 2. Samples of size 4,000 were selected using a complex design. This design was not revealed to the research teams because one would not know such details in practice.

## 4. Evaluating the Known Bias

CCTC has a large set of variables which can be used in evaluating estimation methods. We chose a small subset of variables for research teams to consider when evaluating estimation methods. Evaluation variables were required to have a low percentage of missingness—fewer than 1,135 (1%) missing values in Frame 1. Furthermore, we chose variables related to a person's behavior, and variables related to a person's attitude. Finally, we chose variables that have both high and low know bias. Tables 2a and 2b provide information about the variables chosen.

**Table 2a:** Summary of Attitudinal Evaluation Variables

| Var Name | Quex | Category | Probability Frame Percentage | Nonprobability Frame Percentage |
|---|---|---|---|---|
| q17 | During a crisis like Covid-19, how important or unimportant are arts & culture organizations to you? | Very Unimportant | 0.75 | 1.00 |
| | | Unimportant | 3.59 | 4.16 |
| | | Neither | 14.05 | 15.86 |
| | | Important | 26.07 | 27.44 |
| | | Very Important | 55.54 | 51.53 |
| q18 | Before Covid-19, how important or unimportant were arts & culture organizations to you? | Very Unimportant | 0.16 | 0.32 |
| | | Unimportant | 0.94 | 1.41 |
| | | Neither | 8.27 | 10.94 |
| | | Important | 27.07 | 28.74 |
| | | Very Important | 63.57 | 58.58 |

**Table 3b:** Summary of Behavioural Evaluation Variables

| Var Name | Description | Quex | Probability Frame Percentage | Nonprobability Frame Percentage |
|---|---|---|---|---|
| q7_22 | Classical music | Which of the following activities did you attend or participate in last year | 48.92 | 43.61 |
| q10_1 | Experiencing artworks, performances, or specific | Now that many of those cultural activities are shut down during the | 71.38 | 67.35 |
| q11_1 | Online exhibitions or galleries | Here are some online or digital cultural activities that are being | 57.27 | 53.62 |
| q25_11 | See a play (nonmusical or musical) | Thinking ahead to when people are able to go out again, what are you | 34.81 | 31.73 |
| q1_15 | Participated in a live interactive event online, | Which of the following activities have you done in the past 30 days? | 48.25 | 45.23 |
| q6_9 | Fun | What do you want more of in your life right now? | 45.15 | 46.78 |
| q11_4 | Online materials or activities for kids (for | Here are some online or digital cultural activities that are being | 46.32 | 45.67 |
| q10_3 | Celebrating my cultural heritage | Now that many of those cultural activities are shut down during the | 4.81 | 5.02 |
| q7_14 | Community festival/street fair | Which of the following activities did you attend or participate in last year | 55.19 | 55.08 |
| q6_1 | Hope | What do you want more of in your life right now? | 38.62 | 38.73 |
| q1_6 | Watched a movie or TV series | Which of the following activities have you done in the past 30 days? | 90.87 | 90.79 |
| q25_13 | Take an art, music, or dance class | Thinking ahead to when people are able to go out again, what are you most excited to do in the first few weeks? | 9.43 | 9.50 |

Figure 1a shows the absolute known bias of the chosen attitudinal variables, and Figure 1b shows the absolute known bias of the chosen behavioral variables. The "high bias" evaluation variables have $B_{pop} > .01$ (1 percentage point) and range from 1.4 to 5.3 percentage points.

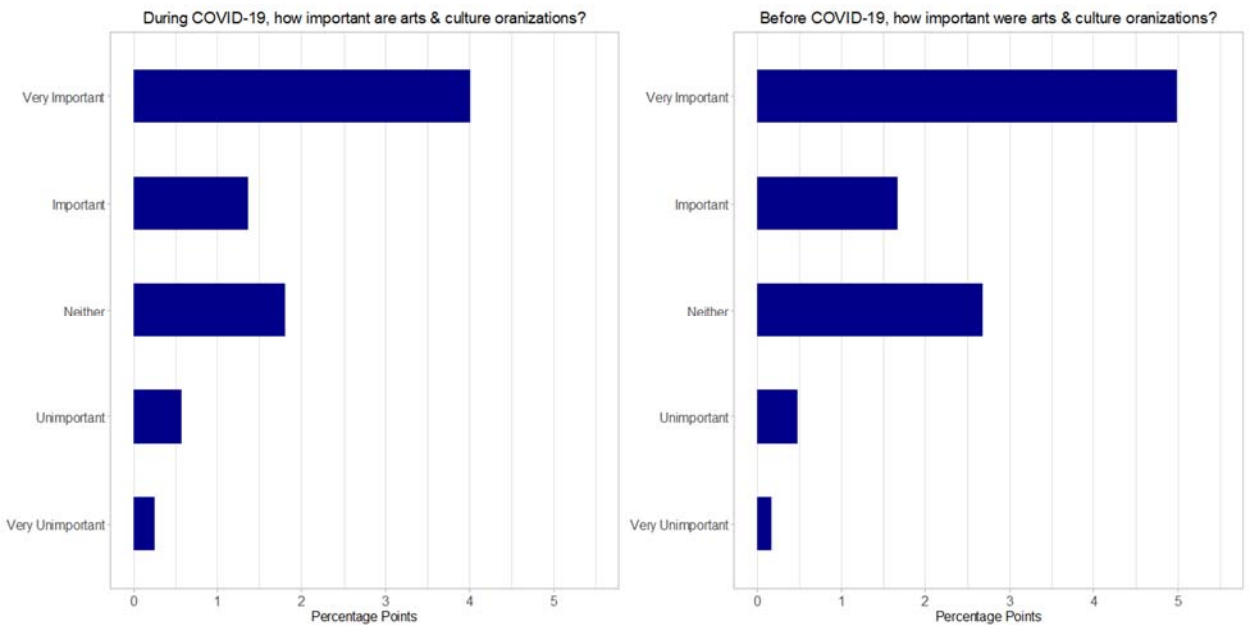


**Figure 1a:** Absolute Known Bias of Attitudinal Evaluation Variables.

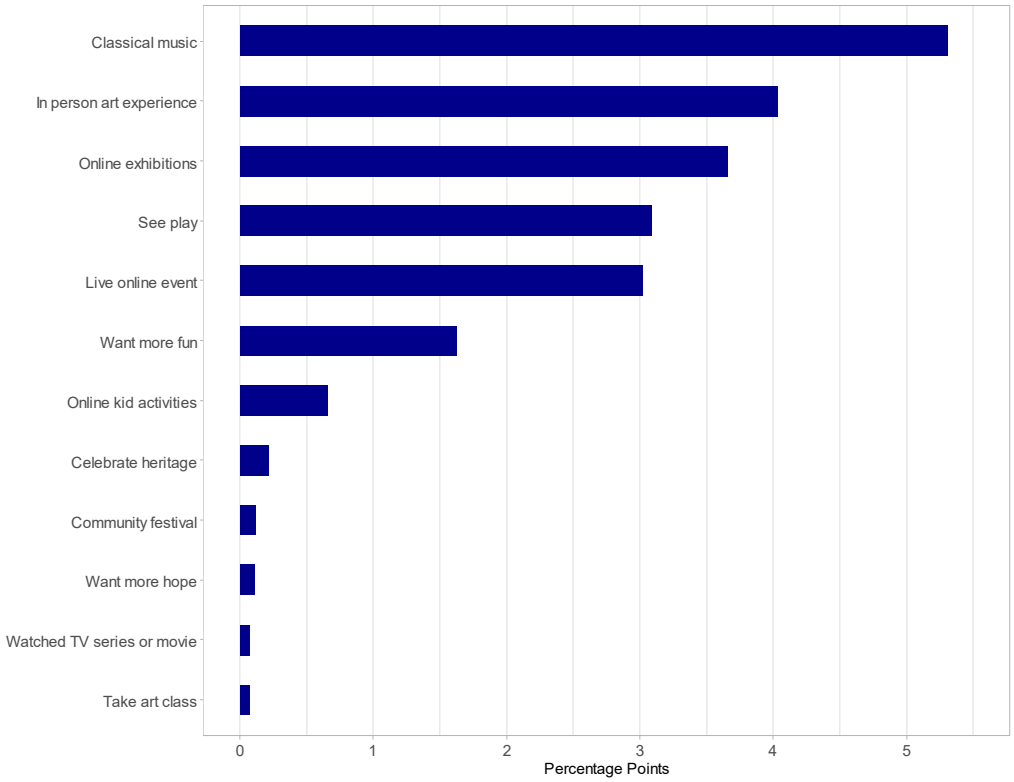


**Figure 1b:** Absolute Known Bias of Behavioural Evaluation Variables.

## 5. Concluding Remarks

Our intentions are to provide a set of probability and nonprobability sample pairs that research teams can use to compare the methods developed for utilizing nonprobability samples, especially in situations where a companion probability sample is available. The simulation design for the nonprobability samples mimics the setting of a common type of nonprobability sample: a sample from an opt-in online panel. Such samples usually come from a subset of the target population, and estimates using these samples will be biased unless attempts are made to address coverage and selection issues.

As with most simulations, we are unable to recreate the real-world complexities that are present in practice. It is also the case that estimation methods that produce successful outcomes from the simulated samples, such as, reduced bias and good confidence interval coverage, may not actually produce such outcomes with real-world samples. However, we believe that the simulated samples constructed for this JSM session will be of use in understanding the models developed for utilizing nonprobability samples, alone or in combination with a probability sample. The strengths and weaknesses of the models will be better understood.

We also hope that there will be more opportunities to create and share additional simulated samples based on different assumptions. The more we can do to test and probe methods for utilizing nonprobability samples, the better we will be able to understand and evaluate estimates based on nonprobability samples.

## References

J. Benoit-Bryan. 2020. “A Time of Crisis: A National Survey of Arts and Culture during COVID-19 with a Focus on Black or African American and Hispanic Voices,” *Journal of Cultural Management and Cultural Policy*, Vol. 2, pp. 49-76. doi 10.14361/zkmm-2020-0203.

R. Valliant. 2020. “Comparing Alternatives for Estimation from Nonprobability Samples,” *Journal of Survey Statistics and Methodology*, Vol. 8, No. 2, pp. 231–263.

Y. M. Yang, N. Ganesh, E. Mulrow, and V. Pineau. 2019. Evaluating Estimation Methods for Combining Probability and Nonprobability Samples through a Simulation Study. In JSM Proceedings, Survey Research Methods Section. Alexandria, VA: American Statistical Association. 1714-1727.