14. Running Surveys for Investigations

Crina-Gabriela Boroş

Abstract

Read this first before running a survey for accountability journalism: dos, don'ts and how to handle imperfect circumstances.

Keywords: statistics, data journalism, surveys, accountability

Is an issue anecdotal or systematic? You're attempting to discover this when you realize there is not any tabular data—a fancy phrase that simply means information supplied in rows and columns. What should you do?

What is data, anyway? There are many nerdy definitions floating around, some of which are intimidating.¹ Let's trade them for the simple concept of "information." And as you gather it, in any shape or form, you need to be able to find patterns and outliers. This means that you have to have a considerable amount of systematically gathered raw material that documents an issue according to a specific method (think fill-in forms). Whether you use a spreadsheet, a coding environment, an app, or pen and paper, it does not matter.

Sometimes, thoughts, feelings or past intimate experiences trapped in people's hearts and minds can be articulated as data. One method of harvesting this precious information is to design a survey that would gather and order such feelings and experiences into a table, archive or a database that nobody else but you has access to.

For instance, the Thomson Reuters Foundation (TRF) undertook a project reporting on how women in the world's largest capitals perceived sexual violence on public transport affects them.² It was a survey-driven effort to

¹ See 130 definitions of data, information and knowledge in Zins, C. (2007). Conceptual approaches for defining data, information, and knowledge. *Journal of the American Society for Information Science and Technology*, 58, 479–493.

² <http://news.trust.org/spotlight/most-dangerous-transport-systems-for-women/?tab=stories>

raise awareness of the issue, but also to compare and contrast (the stuff stats do).

To deliver this spotlight, we went through several circles of Hell, as there are rigorous conventions that social scientific methods, like surveying, require, even when imported by journalists into their practice.

Here are a few main polling rules that journalists would benefit from knowing, but often don't receive training for.

Respondents cannot be handpicked. In order to be considered “representative” a pool of respondents would conventionally include people from all social categories, age groups, education levels and geographical areas that we have to report on. According to established methods, samples of the population under study need to be representative.

The selection of respondents needs to be randomised—meaning everyone has the same chance of having their name drawn from a hat. If you're conducting a poll and speaking to whomever is closest to hand without any criteria or method, there is considered to be a risk of producing misleading data, especially if you are aiming to make more general claims.

The number of people taking a survey must also reach a certain threshold for it to be representative. There are helpful online calculators, like those provided by Raosoft, Survey Monkey or Survey Systems.³ As a rule of thumb: Keep the confidence level at 95% and the margin of error no bigger than 5%.

Answer options must allow respondents to not know or not be certain. When reporters follow these basic rules, their findings are close to unattackable. At the time of the TRF public transport safety research, our polling methodology stuck to the conservative rules of social sciences. Our subject addressed such a common human experience that speaks volumes about how societies function, that a UN agency offered to join in our effort. An honour, but one which, as journalists, we had to decline.

If you like the sound of this, it's time to take a stats course.

Sometimes rigorous polling is unrealistic. This doesn't always mean you shouldn't poll.

While there are established methods for surveying, these don't exhaust what is possible, legitimate or interesting. There may be other ways of doing polls, depending on your concerns, constraints and resources.

For example, when *openDemocracy* wanted to interview reporters across 47 European Council member states about commercial pressure inside newsrooms, there was little chance for statistical significance.

3 <http://www.raosoft.com/samplesize.html>, <https://www.surveymonkey.com/mp/sample-size-calculator/>, <https://www.surveysystem.com/sscalc.htm>

“Why?” you might ask.

All respondents became whistle-blowers. Whistle-blowers need protection, including not disclosing important real demographic data, such as age or sex. We were expecting some contributions from countries where exercising freedom of speech may lead to severe personal consequences. We decided that providing personal data should not be compulsory; nor, if provided, should these data sit on a server with a company that co-owns our information.

The EU had wildly different and incomplete counts of journalists in the region, meaning establishing a country-level representative sample was tricky.

We couldn't line up all press unions and associations and randomize respondents because membership lists are private. They also don't include everyone, although it would have been an acceptable base as long as we were honest about our limitations. Plus, in some countries, transparency projects lead to suppression and we received expert advice in which countries we could not solicit the support of unions without attracting either surveillance or punitive consequences.

In cases like this, you needn't throw the baby out with the bathwater. We didn't.

Instead, we identified what mattered for our reporting and how polling methods could be adjusted to deliver stories.

We decided that our main focus was examples of commercial pressure inside national newsrooms; whether there was a *pattern of how they happened; and whether patterns matched across the region.* We were also interested in the types of entities accused of image-laundering activities in the press.

We went ahead and built a survey, based on background interviews with journalists, media freedom reports and focus group feedback. We included sections for open answers.

We pushed the survey through all vetted journalism organization channels. In essence, we were not randomizing, but we also had no control over who in the press took it. We also had partners—including Reporters sans frontières, the National Union of Journalists and the European Federation of Journalists—who helped spread the questionnaire.

The feedback coming through the survey was added to a unique database, assigning scores to answers and counting respondents per country, drawing comparisons between anecdotal evidence (issues reported sporadically) and systemic issues (problems reported across the board).

The open text fields proved particularly useful: Respondents used them to tip us. We researched their feedback, with an eye for economic censorship

patterns and types of alleged wrongdoers. This informed our subsequent reporting on press freedom.⁴

Although we did publish an overview of the findings, we never released a data breakdown for the simple reason that the selection could not be randomized and country-level sample sizes were not always reached.⁵ But we built a pretty good understanding of how free the press is according to its own staff, how media corruption happens, how it evolves and sadly, how vulnerable reporters and the truth are.⁶

So, are there rules for breaking the rules?

Just a few. Always describe your efforts accurately. If you polled three top economic government advisers on a yes–no question, say so. If you interviewed ten bullying victims, describe how you chose them and why them in particular. Do not label interviews as surveys easily.

If you run a statistically significant study, have the courtesy to release its methodology.⁷ That affords the necessary scrutiny for your audience and experts to trust your reporting. No methodology, no trust.

Don't be the next biggest "fake news" author. If an editor is pushing you to draw correlations based on inferences rather than precise data collection, use a language that does not suggest causality or scientific strength. Our job is to report the truth, not just facts. Do not use facts to cover up a lack of certainty over what the truth is.

Where does your story lie? In a pattern? In an outlier? Decide what data you need to collect based on this answer. Figure out where and how the data can be obtained before you decide on the most appropriate methods. The method is never the point, the story is.

If you run a survey, field-test your findings and protect your reporting against potentially problematic claims. For example, say a survey suggests that the part of the city you live in has the highest crime rate. Yet you feel safe and experienced almost weekly street violence in another neighbourhood you lived in for a year, so you may not yet trust the data. To check if you can trust your data, visit the places that you compare and contrast; talk to people on the streets, in shops, banks, pubs and schools; look at what data was collected; are residents in one area more likely to file complaints than residents in another area? What types of crime are we talking about?

4 <https://www.opendemocracy.net/en/author/crina-boros-2/>

5 <https://www.opendemocracy.net/openmedia/mary-fitzgerald/welcome-to-openmedia>

6 <https://www.opendemocracy.net/en/welcome-to-openmedia/>

7 <http://news.trust.org/spotlight/most-dangerous-transport-systems-for-women/?tab=methodology>

Have the types of crime considered in the analysis been weighted, or does a theft equal a murder? Such “ground truthing” efforts will allow you to evaluate your data and decide to what extent you can trust the results of further analysis.

About the Author

Crina-Gabriela Boros is a UK-based award-winning investigative reporter who teaches data-driven journalism internationally and has established a strategic data journalism training programme at the annual conference Data Harvest.