Extreme Case Sampling on Predictor Variables: A Powerful Qualitative Exploratory Research Method

David Cochrane

Abstract

Qualitative analysis of extreme cases selected from a sample of thousands of data often provide knowledge discovery in a highly efficient, elegant, and powerful manner. This qualitative exploratory research method is frequently overlooked. It is especially useful in cases where (a) a predictor variable is believed to have a linear or non-linear continuously increasing or decreasing effect on an outcome variable, (b) potential predictor variables are well-defined but potential outcome variables are unclear, difficult to measure, or difficult to assign values to, (c) a "double sampling" procedure can be performed where a minimum of 3 cases to a maximum of 5% of cases from a sample of at least 100 subjects that is representative of some population can be obtained, (d) scores for a predictor variable or composite of predictor variables can be assigned to all of the 100 or more subjects from which either the highest or lowest scores can be selected as the cases to study, (e) there are one or more theories or models that can be evaluated by a person or persons with content expertise, (f) these theories or models have not already been confirmed with more definitive quantitative studies, and (g) the possible relationships between predictor variables and outcome variables may be complex and/or there are a large number of potential predictor variables. Even if all seven of these conditions do not exist, studies of the most extreme cases on the predictor variable, whre the extreme cases are selected frm a sample of thousands of data, may still be a very effective research design.

Introduction: Knowledge Discovery in an Era of Big Data

One task of a research methodologist is to select from a very large toolbox of potential research methods the research methods that are most appropriate for a particular question that the researcher has. The research methodologist carefully considers the nature of the question, the specific goals that the researcher has, and the available data and resources that can be applied to answering the question.

We are now in an age where often there is a large amount of data and there are suspected relationships within the data. These suspected relationships within the data may be based on anecdotal evidence or hunches. The researcher wants to explore the possibility that some of these suspected relationships may actually exist. The researcher may point to a few cases where some particular predictor variable may have a relationship to some particular outcome. Precisely what the outcome is and how to mesure the outcome may not be clear.

Consider, for example, this scenario: a graduate student is a believer in astrology. The student's advisor regards astrology as superstition and nonsense but the advisor agrees to explore the possibility of conducting research to determine if relationships in the data can be found.

The student informs the advising professor that:

1. He has about 5,000 birth data from a birth certificate or birth record for famous people and biographies of these people are readily available online.

2. It is believed that the zodiac sign makes Aries pioneering, Taurus stubborn, etc.

3. The student would like to validate that these relationships exist. The advising professor is now taking an interest in this study because he will be happy to see that these supposed relationships are nonsense. Thus, the student and the professor see that a public good can be derived from the research: the student sees the opportunity to open up a new academic field with a pioneering study that confirms an astrological principle, and the professor see the opportunity to provide strong envidence that can help remove the widespread proliferation of superstitious and unhelpful ideas.

The question for the research methodologist is what research design is most effective in building a model to confirm that these relationships exist.

Extreme Case Sampling Defined

Extreme case sampling is simply selecting the most extreme subjects from a large sample on either the predictor variable or the outcome variable. Thus, extreme case sampling may be a double-sampling method where first sample data is obtained from a population and the most extreme cases from this data are used in the analysis. Alternatively, the extreme cases may be selected from the entire population.

Using our scenario of an astrological study, suppose that a person has both Sun and Moon in Aries. This is more Aries than a person who has only Sun or Moon in Aries. By assigning points to a planet in Aries, we can obtain a composite Aries score for each of the 5,000 people in the database. The person with the highest number of points is the most extreme Aries. The individuals who scores highest in Aries provide provide extremely important information that can determine whether the astrological theory is viable, as described below.

Extreme case sampling on the outcome variable is also possible. If we have a way to measure the personality trait of independence, then the person with the greatest independence, then we can select extreme cases based on this expected outcome variable. In the scenario presented here, obtaining scores based on the predictor variable is very simple whereas obtaining scores based on an outcome variable is likely to be much more difficult. Even if a measurement of the outcome variable can be obtained, there are likely to be advantages to sampling on the predictor variable.

A Good, But Imperfect, Sample is Adquate

Let us suppose that in our example scenario that there are about 5,000 birth data to analyze and the data was collected by astrologers in order to have access to data of celebrities, leading politicans, and other famous people. The data was no collected without regard to supporting or defending a theory so obvious selection bias that favors particular astrological beliefs is not a concern. This scenario is typical of existing databases in that the data is very useful but may not be completely beyond any possible bias because it was not collected in a rigorous way that would virtually make bias impossible.

The requirements for exploratory research are not as rigorous as for a hypothesis test. In this exploratory research we are seeking strong evidence that either the astrological hypothesis is valid or not valid. We are not seeking the convincing evidence that a replication of a hypothesis test provides. We are making a large incremental step from anecdotal evidence to a method where most selection bias, and in our sample scenario most likely all seelction bias, is removed by the fact that the researcher does not have the liberty to personally select the data. A database of 5,000 chart data that is available through existing commercial software is used. The scenario that is presented here is typical of many scenarios, such as where databases of student performance, etc. are available in the area of educational psychology. Also, even though our database was not collected using random sampling or systematic sampling, if evidence for the astrological hypothesis is not found through the analysis of extreme Aries individuals, the rejection of the astrologicl hypothesis is very strong because any bias which might exist, even if only remotely possible, would favor the astrological hypothesis.

Thus, if we can demonstrate that Aries does not incline people to be more independent than other people, then we can reject the astrological hypothesis. If certain conditions are met, we can reject (or fail to reject) with good confidence the astrological hypothesis with the available database and without conducting a hypothesis test and without the need for a control group. As we shall see, these conditions are met in the scenario presented in this paper.

An Important Requirement: A Continuously Increasing or Decreasing Graph

Extreme case sampling is a useful research method only if the relationship of the predictor variable and the outcome variable is hypothesized to be continuously increasing or continuously decreasing. This is a critically important requirement and a violation of this requirement results in extreme case sampling being an inappropriate research method.

In our hypothetical scenario, the graduate student informs us that having many planets in the zodiac sign Aries increases the Aries personality trait of independence. It is not critically important whether the relationship is perfectly linear. Perhaps having a fifth planet in Aries as

opposed to four planets in Aries makes very little difference but does slightly increase the inclination to be independent. As long as a line graph with the predictor variable along the x axis and the outcome variable along the y axis is expected to gradually go up (or gradually go down) without any peaks and valleys, then the person with the most Aries is also the person expected to be the most independent.

The hypothesized relationship fits into a regression model where perhaps the weights of the individual components of the composite predictor variable score are unknown but all of these components are added together and none are subtracted. For example, perhaps Sun in Aries is more important than Mercury in Aries so a regression equation would assign a greater coefficient to the Sun sign variable than to the Mercury sign variable. We do not need to know with precision what these coefficients are to derive good information from our extreme case sampling study. We can also create alternative composite predictor variable scores using different guesses for coefficients to test different models. Interactions of these components are also possible as long as the interaction of variables is positive. In practical terms with our scenario, if a person has Sun, Moon and what astrologers call the Ascendant in Aries, the person is an extreme Aries. Virtually all astrologers agree with this. Variations of the formula can be created to test variations of the theory. Thus, we are able to identify with very good confidence individuals who are extreme on the predictor variable. Keep in mind that we are conducting exploratory research to determine if a theory is viable. We are not conducting a rigorous hypothesis test because we do not need to invest the great resources needed to determine if the astrologial hypothesis is viable. As long as content experts (i.e., professional astrologers in this scenario) agree that the selected people based on the given criteria are exremely strong in the trait, we can proceed with the extreme case sampling research procedure to provide strong indications of whether the suspected relationships actually exist.

The power of extreme case sampling derives from the tremendous impact that extreme cases have on a correlation. All statisticians are familiar with the huge impact of extreme cases on a correlation. Note that we should not assume that these extreme cases are outliers. We have almost no assumptions about the distribution of the data and the extreme cases may not be outliers in the sense of being clearly separate from the majority of the data which lies near the middle of the graph. Using the data in our scenario to illustrate the point: if the most extreme Aries people are not independent, we can reject the astrological hypothesis. Because of the enormous impact by extreme cases on a correlation, there would need to be an enormous and extraordinarily unlikely tendency of the rest of the data to have a very strong correlation of Aries with an independent personality. Thus, we can reject the astrological hypothesis by analyzing only a few cases and we can generalize from these cases to an entire population.

Again, keep in mind that this is an exploratory research method that can be conducted with extremely few resources in order to make an enormous incremental step in model building beyond what is achieved with anecdotal evidence. The goal is to develop building blocks beyond

what is possible with even a massive amount of anecdotal evidence by subjecting the hypothesized relationships to a procedure that removes selection bias of the researcher and evaluates the viability of the proposed relationship.

A Summary of Key Points and Additional Important Considerations

What seemed to perhaps be an intractable question because of the enormous resources needed to address the question has become a relatively easy question to answer given the following critically important characteristics of this scenario:

1. There is an available database of data that provides predictor variable data with very low measurement error. In our scenario, this is a database of about 5,000 famous people with recorded birth times and a large amount of biographical information online. A score for Aries (or Taurus, Gemini, etc.) can be computed by adding the number of planets in the zodiac sign and optionally weighting some planets more than others. Content experts (in this case, professional astrologers) are needed to confirm that this formula is reasonable. In the case of our scenario, the extreme people will have so many planets in the zodiac sign that virtually every astrologer will agree that these are extreme cases.

Note that extreme case sampling is a kind of "double sampling" procedure. First, we have some database that may be an entire population but more often is sampled from some population. Then we sample from this sample by assigning scores to each subject in this population based on our predictor variable and we select some number of extreme cases for our study. The number of cases selected for the study is discussed in item #3 below.

2. The relationship of the predictor variable and outcome variable is continuously increasing or decreasing.

3. The number of extreme cases that are selected to analyze should be at least three. In many cases the hypothesized relationship can be rejected after analyzing only the three most extreme cases. For example, we find that the three most extreme Aries individuals do not exhibit independence in any identifiable way more than other people. We then reject the astrological hypothesis or change the astrological hypothesis to fit the findings. We might, for example, proppose that the independent quality is so subtle, psychological or internal that oly a psychological test or interviews can reveal it. Such alternative explanations tremendously erode the original hypothesis. The researcher should set a cut off point for how much data will be analyzed before analyzing the data! Otherwise there is a temptation to cut off after some number of data that fits the hypothesis. The researher should set the number to at least three. The primary determinant of the cut off point is the time and energy that the researcher is willing to commit to the study. In the given scenario where biographies of the famous people need to be studied and sometimes works that they produced analyzed, a researcher with a limited amount of time may wish to study only the top three cases. Studying more than the top 5% (25

in the case of 5,000 data) is not advised because the greater the number of data analyzed, the less extreme some of this data is.

4. The outcome variable may be difficult to assess. Whether a person is independent or not is difficult to determine. In what ways is the person independent? One of the great benefits of extreme case sampling is that we are able to refine and make much clearer what the outcome is! Extreme case sampling assists us in developing a viable model. In my personal experience conducting dozens of extreme case sampling studies, the typical finding is that a more clearly defined outcome variable emerges from the study. The predictor variable is already clearly defined and now the outcome variable becomes much more clearly defined as well. Future studies and anecdotal evidence are much easier to evaluate because we have a much more clearly defined outcome variable rather than an outcome variable that is very general and ambiguous. Extreme case sampling provides an enormous step towards narrowing the scope of possibilities so that future studies can determine whether the scope narrows down to a clear relationship or to the discovery that there is no relationshp of the variables and therefore the hypothesis can be rejected.

5. There may be varous outcome variables that are explored. Extreme case sampling is an exploratory research method that helps us determine what models are viable. We have a choice of whether we want our research to entertain a wide range of possible outcomes and be much more exploratory in nature of whether we have a more specific outcome that we expect. By focusing on a more specific outcome we move more quickly to a confirmation of our hypothesis but we miss the opportunity to notice outcomes that may be a bit different from what we expected. Do we expect Aries to incline towards self-employment, strong personal opinions, aggresssiveness, or any of the above? The wider the net so to speak, the more likely we are to find some relationship but the weaker is our confirmation of a relationship. Regardless of how wide the net is, extreme case sampling enables us to draw much more reliable conclusions about what kinds of relationships may exist between the variables, if any, than through anecdotal evidence gained through convenience sampling. Unlike anecdotal evidence, extreme case sampling also helps develop a body of literature that future researchers can build from.

6. The extreme cases **must** have high scores or ratings on the outcome variable. If they do not, you must reject the the hyothesis. You may radically alter the hypothesis to make it compatibile with the findings but the hypothesis as stated is rejected with very good confidence.

7. Extreme case sampling is particularly helpful when the relationship of predictor variables and outcome variables is complex or there are many competing theories. If a relationship between variables has already been determined by previous research, then a research method that can draw stronger conclusions, like a hypothesis test, may be appropriate. However, when relationships between variables have not yet been clearly determined, hypothesis tests are "shots in the dark" that usually miss the mark. Extreme case sampling can sort out a tangle of data into promising connections and also remove from consideration many

relationships. If Aries has no relationship to independence, then it is extremely unlikely that an independent trait will be found in the top 3 or more extreme Aries people. At best, the researcher will need to narrow done to some particular definition of independence that is tremendously more specific and easier to test than the much more general notion of independent personality that was initially proposed.

Advantages of Sampling on the Predictor Variable

By sampling on the predictor variable, rather than the outcome variable, we are not limited to narrowly defined outcomes. We can consider a wide range of possible outcomes. The potential for knowledge discovery is enhance as we explore possible relationships. The extreme cases based on the predictor variable are believed to have some effect but the exact nature of the effect does not need to be known and unexpected relationships may be discovered. By qualitatively analyzing these extreme cases we have also opened the door to discovering relationships that are not easily quantified. Thus, this research procedure has cast the widest possible net for finding any possible impact from the predictor variable. If no viable outcomes can be proposed from these extreme cases, then we have obtained very strong evidence that the belief that these variables are predictors of *any* kind of outcome is very strong.

A failure to find any outcomes associated with strong Aries in our hypothetical scenario suggests very strongly that the belief in Aries is superstitious. This conclusion is much stronger than a negative finding from other froms of research which may fail because (a) there are outcomes that were not investigated, (b) the measurement of the outcome suffered from too much measurement error to detect the outcome sufficiently, or any other problems that might arise from the limitations of selecting particular outcomes and atempting to measure them.

Success in finding some traits or behaviors that are shared among these thexe extreme cases provides the researchers with ideas, hypotheses, or theories that have the potential to be verified in future research.

Extreme Case Studies that Do Not Use Extreme Case Sampling

Not all extreme case studies use extreme case sampling. For example, an astrologer may have a client who has several planets in Aries. Conclusions derived from the analysis of these anecdotal extreme cases are much weaker than conclusions derived from extreme cases that are sampled from a database. The extreme cases that are anecdotal cases can very easily be influenced by selection bias. The astrology may, for example, may be more responsive to cases that confirm a belief and consciously or unconsciously avoid analyzing extreme Aries cases that contradict the astrological belief. Therefore, the term "extreme case sampling" is preferred to "extreme case study" in describing this methodology because not all extreme cas studies meet the criteria for being able to draw strong evidence from the study.

Conclusion

There are particular research questions that extreme case sampling is remarkably well suited for. However, extreme case sampling often gets ignored as an important research method. For example, belief in astrology is pervasive with most surveys indicating that about 1/3 of the population of many countries believing in astrology. Those who believe that astrology is a harmful superstition could use extreme case sampling to demonstrate that the ideas are not valid, and the believers could use extreme case sampling to build viable models. The resources needed for extreme case sampling are readily available in the area of astrology and in many other areas. The scnario presented in this paper is is only one example of a situation where extreme case sampling is a brilliant tool in our toolkit of research methods. Extreme case sampling is often overlooked, and researchers in many fields are likely to find that extreme case sampling needs to be resurrected as an invaluable reserach method that has gotten buried beneath the other valuable tooks in our research methods toolkit.