Section 2: Week 3: Designing an Experiment

Nate Bachmeier

TIM-7101: Statistics for Technology Leaders

August 16, 2020

North Central University

# Designing an Experiment

NCU-Cares (NCU-C) is a politically neutral nonprofit organization, seeking to make the world a better place through targeted lobbying efforts. The death of George Floyd has risen the debate of police violence and reform to the national stage (Crary & Morrison, 2020). While the topic rests on American’s hearts and minds, it has also become highly partisan with many efforts to undermind the conversation (McCaskill, 2020). Separating the seed from the chaff requires a well-structured experiment that identifies independent and dependent variables (IV versus DV). Next, precise controls must exist to reduce interference from either internal or external threats to conclusive validity. These controls will guide data collection requirements and dictate analysis procedures that follow.

# Section I: Identify the Problem

The first step to designing an experiment is to define clearly the specific problem of research. Harris (2008) uses an example of investigating the “influence of music on driving ability.” In this situation, music is the independent variable versus driving ability the dependent result. However, music could refer to either the categorical genre or perhaps the volume level. Meanwhile, the driver’s ability could be a measurement of maintaining lane alignment or parallel parking. These nuances to the question have a substantial impact on all aspects of the design and must be declared upfront.

Similarly, an exploration of police violence needs explicit constraints to avoid boiling the ocean through open-ended discovery. The collective perspective of the Black Lives Matter movement is that police exert disproportionate force against people of color (BLM, 2020). This statement infers that *race* is the independent variable that shapes the *response*. Meanwhile, others believe that *sanity*, such as caused by mental illness or narcotics, applies more pressure to the situation (Lamb, Weinberger, & DeCuir, 2014). NCU-C would like to understand which of these independent variables is the most dominant factor. After assessing this information, the organization could then more effectively fight police violence through either (a) civil rights policies or (b) mental health programs. However, a null hypothesis could exist that concludes that there is no statistically significant difference between these variables. That result might infer that law enforcement indiscriminately applies force to all parties.

# Section II: Methodology

There are multiple strategies for determining which variable has more effect on a situational outcome. News articles typically approach the problem by looking at the raw descriptive statistics, such as the ratio of victims that were experiencing a mental crisis. A challenge with this solution is that the telemetry only communicates what happened, not why. Consider the extreme example that one hundred percent of all police violence within a given community is against a specific race. While this scenario immediately raises questions around racial profiling, it should also invite a discussion around the diversity of the inflicted population. An alternative solution could look at changes after significant interventions (DeCarlo, 2018). Starting in the late 1960s, health institutions began releasing and turning away thousands of patients due to insufficient funding (Lyons, 1984). The impact of these decisions has likely left an imprint in arrest policies and statistics. If such an imprint does exist, then examining funding records on mental health and drug addiction facilities might surface a correlation.

DeCarlo (2018) states that quasi-experiments are particularly useful in social welfare policy research (see chapter 12.2). Under a quasi-experiment, the researcher team does not use random assignment and instead looks at different populations. This method could be highly effective for examining the impact of both *race* and *sanity* variables. For example, how does *race* impact police violence when comparing diverse metropolitan areas (e.g., Chicago and Detroit) to homogenous cities (e.g., Brandsen or Sioux Falls)? Likewise, for every dollar that Nevada spends on public health, Alaska invests six (United Health Foundation, 2017). From examining these groups that are both similar and complete opposites, it should lead to a quantitative sense of the underlying effect of these variables.

# Section III: Threats to Validity

There are four major categories of threats to the accuracy of statistical conclusions that led to erroneous results that are not generalizable (Parker, 1993). When designing high-quality experiments, the designers need to be cognizant of these issues and their sources (see Table 1). Fundamentally, these challenges represent a degradation of the confidentiality, integrity, and availability of the experiment. These limitations prevent the generalization and reproducibility of research, resulting in the discrediting of publications and professional embarrassment (García-Pérez, 2012). Instead, researchers must decide on controls and procedures before even beginning data collection.

Table 1: Threat Sources

|  |  |
| --- | --- |
| Source | Description |
| Internal Threat | Contamination by the research team |
| External Threat | Contamination outside of the study’s controls |
| Statistical Conclusion Validity | Results are arbitrary or non-reproducible |
| Construct Validity | Controls are not enforceable or consistent |

## How do these threats manifest

For example, if a participant needs to provide personally sensitive information, then the data collection must convey trustworthiness (construct validity). Otherwise, the candidate is likely to hold back data like side-effects that are highly relevant to the research project. Without trust, the contributor might engage in activities that directly go at odds with the study (external threats). The research team might lack a plan to deal with these confounding variables, and deviate arbitrarily exclude group members (internal threats). Since these results are now arbitrary, shoehorning outcomes into various statistical models eventually lines up, and the paper is ready for publishing (statistical conclusion validity). Virtually an unlimited number of these permutations exist, and it is impossible to remove all of them (Parker, 1993). However, any procedure that reduces the influence of garbage-in/garbage-out experimentation is ideal.

# Section IV: Data Requirements

For an experiment to be successful, it needs to have sufficient *power* to measure the *effect* in question. Several knobs feed into the power of an experiment, such relaxing the confidence interval, using parametric statistics, converting to a one-tail model, increasing the samples, or adjusting the sensitivity (Donovan, 2016). Deciding which value to tweak is scenario-specific, and can be somewhat of an art form.

## Determining Power

There are fifty-three million people who have an interaction with the police each year (BJS, 2015). Of these encounters, approximately one thousand concludes with police brutality (KBP, 2020). These figures suggest that roughly 0.0019% of all interactions end with the officer killing the suspect. Given the relatively small sample count, it might be necessary to adjust the confidence intervals to meet acceptable power requirements. Another option might be to reduce the number of racial categories, say from nationalities to three groups. These data tweaks might enable seeing high-level trends that future research could tease further.

## Determining Effect

Effect size measures the strength of a phenomenon (Donovan, 2016). While calculating the difference between the two distributions is relatively straight forward, it can be difficult to predict ahead of time. This bittersweet relationship introduces challenges when determining the appropriate sample size. One potential solution is to use an iterative sequential sampling policy instead of a fixed size upfront (García-Pérez, 2012). In this situation, that would mean first choosing two similar populations (e.g., states) and comparing the police violence by *race* and *sanity* as independent variables. While this small group would have a reasonably low confidence interval, it could qualitatively hint at the overall sample size needing to be small, medium, or large. There are potential risks that the random-initial sample produces an invalid seed into the study, but the alternative is to roll a pair of dice or flip a lucky coin.

## Potential Sample Sizes

Despite the effect size being unknown potential, it is possible to determine the range of sample sizes for the experiment (see Table 2). G\*Power version 3.1.9.7 projects that t-tests of the “difference between two independent means (two groups)” for a one-tail model will need somewhere from 4 to 1580 examples. Since public data sets of police violence are generally under 4500 examples, there should be sufficient coverage assuming the specific measurements are kept simple.

Table 2: Sample Sizes

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Power | Effect Size | Confidence – 50% | Confidence – 80% | Confidence – 95% | Confidence – 99% |
| 70% Adequate | 0.20 – Small | 28 | 188 | 472 | 816 |
| 0.50 – Medium | 6 | 32 | 78 | 134 |
| 0.80 – Large | 4 | 14 | 32 | 54 |
| 95% Excellent | 0.20 – Small | 272 | 620 | 1084 | 1580 |
| 0.50 – Medium | 44 | 100 | 176 | 256 |
| 0.80 – Large | 18 | 40 | 70 | 102 |

# References

BJS. (2015, October). *Contacts between police and the public*. Retrieved from Bureau of Justice Statistics: https://www.bjs.gov/content/pub/pdf/cpp15\_sum.pdf

BLM. (2020, May 30). *#DefundThePolice*. Retrieved from Blac Lives Matter: https://blacklivesmatter.com/defundthepolice/

Crary, D., & Morrison, A. (2020, June 11). *Black Lives Matter goes mainstream after Floyd's death*. Retrieved from AP News: https://apnews.com/347ceac3ea08978358c8c05a0d9ec37c

DeCarlo, M. (2018). *Principles of Sociological Inquiry: Qualitative and Quantitative Methods.* Blackstone Textbooks. Retrieved from https://scientificinquiryinsocialwork.pressbooks.com/chapter/12-1-experimental-design-what-is-it-and-when-should-it-be-used/

Donovan, C. (2016, August 30). *Power and effect size*. Retrieved from YouTube: https://www.youtube.com/watch?v=9LVD9oLg1A0

García-Pérez, M. A. (2012). Statistical conclusion validity. *Frontiers in Psychology, 3*. doi:https://doi.org/10.3389/fpsyg.2012.00325

Harris, P. (2008). *Designing and reporting experiments in psychology.* New York, New York, USA: McGraw-Hill Education. Retrieved from https://ebookcentral.proquest.com/lib/ncent-ebooks/detail.action?docID=345135

KBP. (2020, August 11). *Police shootings database*. Retrieved from Killed By Police: https://killedbypolice.net/

Lamb, H., Weinberger, L., & DeCuir, W. (2014). The police and mental health. *Psychiatric Services, 53*(10), 1266-1271. doi:https://doi.org/10.1176/appi.ps.53.10.1266

Lyons, R. (1984, October 30). *How the release of mental-patients began*. Retrieved from The New York Times: https://www.nytimes.com/1984/10/30/science/how-release-of-mental-patients-began.html

McCaskill, N. (2020, July 16). *Trump has tried to make 'Black Lives Matter' a partisan issue. A federal agency says it is not.* Retrieved from Politico: https://www.politico.com/news/2020/07/16/trump-special-counsel-black-lives-matter-366177

Parker, R. (1993). Threats to the validity of research. *Rehabilitation Counseling Bulletin, 36*(3), 130-138. Retrieved from https://search-ebscohost-com.proxy1.ncu.edu/login.aspx?direct=true&db=eric&AN=EJ458938&site=eds-live

United Health Foundation. (2017). *Public health funding*. Retrieved August 16, 2020, from America's health rankings: https://www.americashealthrankings.org/explore/annual/measure/PH\_funding/state/ALL