SCIENTIFIC WHITEPAPER:

STATISTICAL CONSIDERATIONS IN THE DESIGN AND ANALYSIS OF SARS-COV-2 PREVALENCE STUDIES

UNC GILLINGS COVID-19 DASHBOARD

gillingscovid19.unc.edu

Author: Bonnie E. Shook-Sa, Department of Biostatistics, UNC-Chapel Hill

Thanks to:
Jessie K. Edwards, Penny Gordon-Larsen, Allison Aiello, and Ross Boyce.

Prevalence studies play a crucial role in responding to the COVID-19 pandemic. Estimating the number of individuals who have been infected with SARS-CoV-2 by the number of positive clinical tests is likely to be an underestimate due to shortages of test supplies, other barriers to testing, and the high proportion of subclinical infections that do not prompt testing. Point prevalence and seroprevalence studies can offer better estimates of the true number or proportion of infections within a given population. Additionally, these studies can be used to model transmission dynamics, identify risk factors for infection, and compare disease prevalence across population subgroups and/or over time.

This document summarizes some of the statistical issues that should be considered when designing a SARS-CoV-2 prevalence study. A careful design is needed to allow the results of the study to generalize to the target population. Because of the statistical issues involved, it is best to consult with a statistician or epidemiologist at the design phase. In North Carolina, several universities across the state have statistics, biostatistics, and/or epidemiology departments (for example, UNC-Chapel Hill, Duke University, NC State University, East Carolina University), and contract research organizations specializing in sampling methods are another potential resource (for example, Westat, RTI International, Abt Associates).

Design by AdrialDesigns.com

1. TYPES OF PREVALENCE STUDIES

1.1 POINT PREVALENCE

Point prevalence studies aim to estimate the number or proportion of active SARS-CoV-2 infections in a population. These studies provide researchers with data to model transmission dynamics and to evaluate risk factors for infection. Active infections are those in which SARS-CoV-2 RNA is detectable from a nasal or throat swab using reverse transcription-polymerase chain reaction (RT-PCR) tests. The sensitivity and specificity of PCR tests vary by the type of test, disease severity, and the timing of testing relative to symptom onset.

REFERENCES/ADDITIONAL RESOURCES:

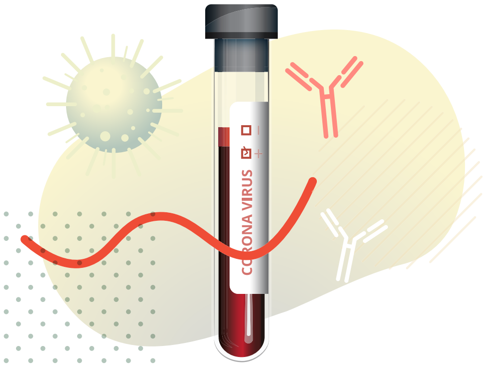
- Centers for Disease Control and Prevention. Interim Clinical Guidance for Management of Patients with

Confirmed Coronavirus Disease (COVID-19). Available at: https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-guidance-management-patients.html

- Kucirka, L.M., Lauer, S.A., Laeyendecker, O., Boon, D. and Lessler, J., 2020. Variation in false-negative rate of reverse transcriptase polymerase chain reaction–based SARS-CoV-2 tests by time since exposure. Annals of Internal Medicine. Available at: https://www.acpjournals.org/doi/10.7326/M20-1495

1.2 SEROPREVALENCE

Antibodies are proteins that the body makes in response to an infection. In the case of COVID-19, measurement of antibodies against SARS-CoV-2 can tell us something about whether a person has been previously infected with the virus. Seroprevalence studies aim to estimate the proportion of the target population with detectable SARS-CoV-2 antibodies. Serological (antibody) tests identify the presence of SARS-CoV-2 antibodies in the blood, and positive tests are considered evidence of a prior infection. People infected with SARS-CoV-2 typically develop antibodies 1-3 weeks following infection, though for some



Design by AdrialDesigns.com

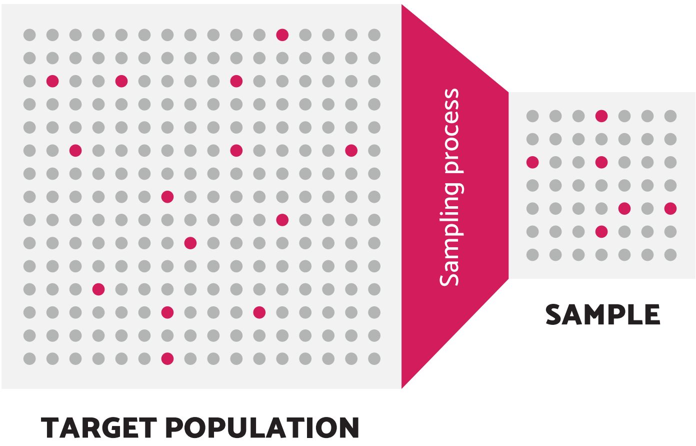
individuals antibodies take longer to develop and some individuals, particularly those with immune compromise or mild disease, may not develop detectable levels of antibodies. Further, it is not yet known how long antibodies can be detected following infection and how long protection against subsequent infection may last. Serological tests have an important role in surveillance because they aim to detect antibody levels to prior infection, including among individuals who were asymptomatic (did not show signs of the disease) or did not seek care (either due to having a less severe case or due to barriers to care).

REFERENCES/ADDITIONAL RESOURCES:

- Centers for Disease Control and Prevention. What COVID-19 Seroprevalence Surveys Can Tell Us. Available at: https://www.cdc.gov/coronavirus/2019-ncov/covid-data/seroprevalance-surveys-tell-us.html
- Centers for Disease Control and Prevention. COVID-19 Serology Surveillance Strategy. Available at: https://www.cdc.gov/coronavirus/2019-ncov/covid-data/serology-surveillance/index.html
- Clapham, H., Hay, J., Routledge, I., Takahashi, S., Choisy, M., Cummings, D., Grenfell, B., Metcalf, C.J.E., Mina, M., Barraquer, I.R. and Salje, H., 2020. Seroepidemiologic Study Designs for Determining SARS-COV-2 Transmission and Immunity. Emerging Infectious Diseases, 26(9).
- Gronvall G, Connell N, Kobokovich A, West R, Warmbrod KL, Shearer MP, Mullen L, and Inglesby T, 2020. Developing a national strategy for serology (antibody testing) in the United States. The Johns Hopkins Center for Health Security. Available at: https://www.centerforhealthsecurity.org/our-work/pubs_archive/pubs-pdfs/2020/200422-national-strategy-serology.pdf

Design by AdrialDesigns.com

- Winter AK, Hegde ST. The important role of serology for COVID-19 control. The Lancet Infectious Diseases 2020; 20(7): 758–759.



2. TARGET POPULATION AND SAMPLING

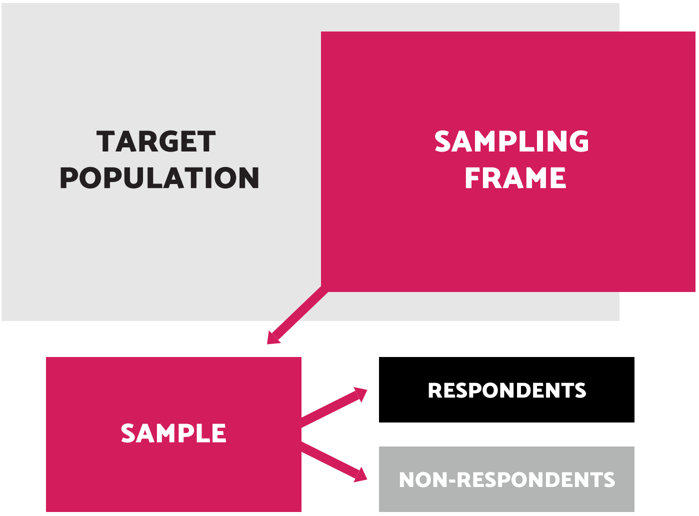
The target population is the population to which the research team wants to make inference. For a SARS-CoV-2 prevalence study, the target population could be the general population, meaning all persons who reside in a given country, state, county, or municipality. Alternatively, the target population could consist of a subset of the general population, such as hospitalized patients, pregnant women, or persons in a school system.

If the goal is to estimate point prevalence or seroprevalence of SARS-CoV-2 within a target population, sampling is often conducted. This is because it is typically too time and resource-prohibitive to test everyone in the target population. If appropriate sampling methods and analytic techniques are followed, unbiased estimators can be used to obtain estimates of point prevalence or seroprevalence from a relatively small sample of the population. The generalizability of the sample results back to the target population depends on the sampling methods, analytic approaches, and the validity of any underlying assumptions on which these methods rely. Regardless of the sampling method used, the sample should be a subset of the target population.

2.1 SAMPLING METHODS

There are two broad categories of sampling methods for prevalence studies. When the sample is selected from the target population using a random process, this is known as probability-based sampling. When selection into the sample is not random, this is known as nonprobability sampling. There are biases that can be introduced in either setting, and the analytic approaches employed must consider these potential biases and adjust for them to the extent possible. The sections below provide more details about each sampling approach and provide examples of biases that can be introduced with these designs.

2.1.1 PROBABILITY-BASED SAMPLING



With probability-based sampling, selection into the sample is random. The researcher develops a sampling frame, which is a list of members of the target population. From the sampling frame, the researcher

Design by AdrialDesigns.com

randomly selects a sample of persons or households and recruits them for participation in the prevalence study. Because selection into the sample is random and not driven by the participant or the researcher, when survey weights are applied (as discussed below), the sample is representative of the target population in expectation.

With probability-based sampling, each unit i has a known probability of selection, π_i . Researchers have the choice of whether to select units from the sampling frame with equal probability (i.e., $\pi_i = \pi_j$ for all i, j) or whether to select some units with higher probability than others (i.e., $\pi_i = \pi_j$ for some i, j). When some units are selected with higher probability compared to others, this is known as oversampling. Oversampling is typically implemented using stratification, where units on the sampling frame are divided into mutually-exclusive and exhaustive strata, and separate random samples are selected within each stratum. In a prevalence study, stratification can facilitate oversampling of population subgroups (for example, populations more vulnerable to COVID-19).

Cluster sampling is another common tool used with probability-based designs. Researchers again divide the sampling frame into mutually-exclusive and exhaustive groups (called clusters). Instead of sampling within all groups, the researcher randomly selects a subset of clusters for inclusion in the study. Cluster sampling helps facilitate logistic feasibility when (1) data collection will take place in person and thus sampled units need to be more closely grouped geographically, and/or (2) a sampling frame of the entire population is not available, so the researcher selects clusters from which sampling frames can later be obtained. Clusters can be selected with equal probability or selected using probability proportional to size (PPS) sampling. With PPS sampling, clusters are selected proportional to a size measure.

Probability sampling tends to take more time to implement and be more expensive than nonprobability sampling, because it takes time

Design by AdrialDesigns.com

to design the sample and recruit participants rather than relying on participants to self-select into the study sample. To obtain estimates more quickly with probability-based sampling, researchers can partner with an ongoing representative cohort within the geography of interest and can recruit participants for the seroprevalence study within that cohort. For example, researchers at UNC are recruiting participants for the Chatham County COVID-19 Cohort (C4) seroprevalence study from the ongoing Chatham Community Assessment, an ongoing representative panel of Chatham County residents. Partnering with an ongoing study reduces the time and cost associated with selecting a new probability sample and facilitates recruitment because participants in the cohort have already participated in the partner study. Of course, this approach is only viable in areas with existing probability-based studies.

SAMPLING FRAMES

When selecting a sampling frame for a probability-based study, it is critical to consider what the target population is for that study and how recruitment will be conducted. For studies of the general population, address-based sampling (ABS) frames, enumerated lists, and Random Digit Dialing (RDD) frames are commonly used. Custom sampling frames are used for target populations consisting of subsets of the general population. The ideal sampling frame includes as many members of the target population as possible, with few members who are outside the target population. It is necessary to consider when the sampling frame was constructed, and which members of the target population might be excluded from the frame.

- **Address-based Sampling (ABS):** ABS sampling frames are based on commercially-available versions of the United States Postal Service’s mailing address database. They have been validated in numerous settings to provide high coverage of the general population, and facilitate recruitment by mail or in-person. Phone numbers can be appended to facilitate limited contacts by phone. ABS frames are frequently updated, and lists can be purchased from address vendors relatively inexpensively.

- **Enumerated Lists:** Enumerated frames are “build-your-own” sampling frames, where researchers sample a geographic area and then canvas the area to enumerate households. These frames are typically more time consuming and costly than ABS frames, but can be used in areas where ABS coverage is poor.
- **Random Digit Dialing (RDD):** RDD frames have been used for general population surveys since the 1980s. They are used for telephone recruitment, allowing sampling of both cell phones and landlines. Careful consideration must be given when mapping RDD frames to geographic areas at the sub-national level.
- **Custom Frames:** Custom sampling frames are common when the target population is a subset of the general population. For example, for a seroprevalence study at a university, the sampling frame might be comprised of all faculty, students, and staff affiliated with the university. For a survey of healthcare workers in a given area, the sampling frame might include lists of physicians, nurses, and healthcare specialists working in healthcare facilities in that area.

POTENTIAL SOURCES OF BIAS

Because selection into the sample is random, selection bias due to the sampling process is eliminated for probability-based samples. However, other errors can be introduced.

Coverage error: Coverage error occurs when members of the target population are excluded from the sampling frame. The level of bias incurred depends on the proportion of the population missing from the sampling frame and how much their prevalence differs from the population included on the sampling frame. For example, there is interest in estimating seroprevalence among persons 65+ but nursing homes and assisted living facilities are excluded from the sampling frame, this could lead to a misleading estimate of seroprevalence. Coverage error can be minimized with careful selection of an appropriate sampling frame and use of supplemental sampling frames, as needed.

Nonresponse error: Nonresponse error can occur when some members of the sample do not participate in the study. The level of bias incurred depends on the proportion of the sample who refuse to participate and how much their prevalence differs from participating individuals. Nonresponse error can be minimized with appropriate study design, including the use of incentives, multiple contacts, and community engagement efforts.

ANALYTIC APPROACHES

Probability-based samples are commonly analyzed using methods from the finite-population inferential paradigm. Sampling weights w_i are assigned to each member of the sample, typically equal to the reciprocal of their probability of selection, (i.e., $w_i = \pi_i^{-1}$. Sampling weights indicate the number of members of the target population represented by each sampled individual. Sampling weights can be further adjusted to account for sampling frame undercoverage and/or nonresponse using raking techniques or calibration estimators. Common statistical software packages, including R, SAS, Stata, SPSS, and SUDAAN, have built-in procedures to appropriately analyze survey data, accounting for the features of the design such as weighting, stratification, and/or clustering.

REFERENCES/ADDITIONAL RESOURCES:

- Frasier A, Guyer H, DiGrande L, Domanico R, Cooney, D, Eckman, S. Design for a Mail Survey to Determine Prevalence of SARS-CoV-2 Antibodies in the United States. Survey Research Methods 2020; 14(2). Available at: https://ojs.ub.uni-konstanz.de/srm/article/view/7757
- Groves RM, Fowler Jr FJ, Couper MP, Lepkowski JM, Singer E, Tourangeau R. Survey Methodology: John Wiley & Sons, 2009.
- Harter, R., Battaglia, M.P., Buskirk, T.D., Dillman, D.A., English, N., Fahimi, M., Frankel, M.R., Kennel, T., McMichael, J.P., McPhee, C.B. and DeMatteis, J.M., 2016. Address-Based Sampling. Prepared for AAPOR Council by the Task Force on Address-based sampling, Operating Under the Auspices of the AAPOR Standards Committee. Available at: https://www.aapor.org/Education-Resources/Reports/Address-based-Sampling.aspx
- Heeringa SG, West BT, and Berglund PA (2017). Applied Survey Data Analysis. CRC press.
- Moraes Silva, D. R., & Mont'Alverne, C. (2020). Identifying impacts of COVID-19 pandemic on vulnerable populations. Survey Research Methods, 14(2), 141-145. https://doi.org/10.18148/srm/2020.v14i2.7742
- Post, J., Class, F., & Kohler, U. (2020). Unit nonresponse biases in estimates of SARS-CoV-2 prevalence. Survey Research Methods, 14(2), 115-121. https://doi.org/10.18148/srm/2020.v14i2.7755
- Schnell R, Smid M. Methodological Problems and Solutions for Sampling in Epidemiological SARS-CoV-2 Research. Survey Research Methods 2020; 14(2). Available at: https://ojs.ub.uni-konstanz.de/srm/article/view/7749
- Shook-Sa BE, Boyce RM, and Aiello AE (2020), Estimation without Representation: Early SARS-CoV-2 Seroprevalence

Design by AdrialDesigns.com

Studies and the Path Forward, The Journal of Infectious Diseases, https://doi.org/10.1093/infdis/jiaa429

- Stringhini S, Wisniak A, Piumatti G, et al. Seroprevalence of anti-SARS-CoV-2 IgG antibodies in Geneva, Switzerland (SEROCoV-POP): a population-based study. The Lancet 2020. Available at: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)31304-0/fulltext

2.1.2 NONPROBABILITY SAMPLING

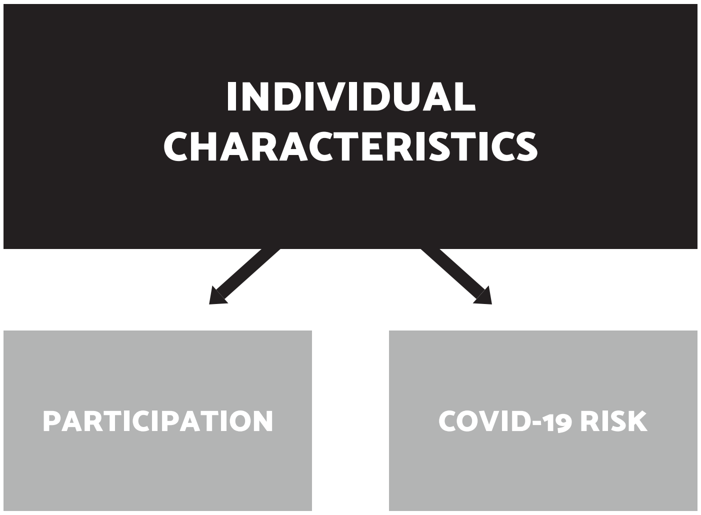
With nonprobability sampling, often referred to as convenience sampling, selection from the target population into the sample is not random. Instead, samples are selected from "convenient" populations or using subjective methods. Because selection is not random, the characteristics of nonprobability samples often differ from those of the target population. Researchers using nonprobability designs must rely on analytic adjustments with estimators that are unbiased under a given set of assumptions. When these assumptions do not hold, estimates from nonprobability samples do not generalize to the target population.

SAMPLING FRAMES

Several SARS-CoV-2 seroprevalence studies have been conducted to date using nonprobability sampling. Examples of nonprobability sampling frames are volunteers recruited on social media platforms, patient populations, or individuals at shopping centers.

Design by AdrialDesigns.com

POTENTIAL SOURCES OF BIAS



Selection Bias: Selection bias occurs when some parts of the target population are not included in the sample, or when some members of the target population are sampled at different rates than intended by the researcher.

Because selection into a nonprobability sample is not random, there can be characteristics of individuals that are associated both with their chances of being included in the nonprobability sample and also with their risk for SARS-CoV-2 infection. For example, individuals with suspected prior COVID-19 infections due to recent symptoms might be more likely to participate in a seroprevalence study compared to asymptomatic individuals. Certain demographic groups (For example age groups or race/ethnic groups) might be more or less likely compared to other groups to volunteer for a prevalence study due to the recruitment method(s) used.

ANALYTIC APPROACHES

Because of the likely differences between the nonprobability sample and the target population, analytic adjustments are needed to

Design by AdrialDesigns.com

generalize the results of the nonprobability sample to the target population. Weighting, model-based, and doubly-robust estimators seek to adjust for the characteristics that affect participation and/or infection risk. These methods rely on the assumption that the nonprobability sample is like a stratified random sample from the target population, where the adjustment characteristics define the strata. This assumption would not hold in a setting where the data are missing not at random (i.e., when participation in the study is directly driven by COVID-19 infection status). This assumption is also violated when an important factor drives both study participation and infection risk (for example, age, race/ethnicity), but is not controlled for in the analysis. Unfortunately, this assumption cannot be validated. These adjustment factors should be identified based on subject-matter expertise when planning the study to ensure collection of these characteristics during data collection.

REFERENCES/ADDITIONAL RESOURCES:

- Baker R, Brick JM, Bates NA, Battaglia M, Couper MP, Dever JA, File KJ, and Tourangeau R (2013). Report of the AAPOR task force on non-probability sampling. Available at: https://www.aapor.org/AAPOR_Main/media/MainSiteFiles/NPS_TF_Report_Final_7_revised_FNL_6_22_13.pdf
- Havers FP, Reed C, Lim T, et al. (2020) Seroprevalence of Antibodies to SARS-CoV-2 in 10 Sites in the United States, March 23-May 12, 2020. JAMA Intern Med. Published online July 21, 2020. doi:10.1001/jamainternmed.2020.4130. Available at: https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/2768834
- Lavrakas PJ (2008). “Nonprobability Sampling.” Encyclopedia of Survey Research Methods. Thousand Oaks, CA: Sage Publications, Inc. doi: 10.4135/9781412963947

Design by AdrialDesigns.com

- Lohr SL (2009). Sampling: Design and Analysis. Nelson Education.
- Valliant, R (2020). Comparing Alternatives for Estimation from Nonprobability Samples, Journal of Survey Statistics and Methodology, Volume 8, Issue 2, Pages 231–263, https://doi.org/10.1093/jssam/smz003

3. ADJUSTING FOR TEST SENSITIVITY AND SPECIFICITY

Prevalence studies rely on the results of imperfect diagnostic tests. Ignoring the sensitivity and specificity of the test used can lead to biased prevalence estimates. Sensitivity is the probability that an individual tests positive given that s/he is infected, while specificity is the probability that an individual tests negative given that s/he is not infected. Estimates from prevalence studies can be adjusted to account for the sensitivity and specificity of the PCR or serology test used.

Rogan Gladen proposed the following estimator of the true population prevalence p:

$$\hat{p} = \frac{\hat{t} + S_p - 1}{S_e + S_p - 1}$$

where $\hat{t}$ represents the sample estimated proportion of persons testing positive, S_e represents the sensitivity of the test, and S_p represents the specificity of the test. Consideration needs to be given to whether sensitivity and specificity are known with certainty or estimated. Estimated standard errors and confidence intervals should also account for test sensitivity and specificity and, when applicable, the fact that these quantities were estimated rather than known.

REFERENCES/ADDITIONAL RESOURCES:

- Lang Z and Reiczigel J (2014). Confidence limits for prevalence of disease adjusted for estimated sensitivity and specificity. Preventive veterinary medicine, 113(1), pp.13-22.
- Mfueni E, Devleesschauwer B, Rosas-Aguirre A, Van Malderen C, Brandt PT, Ogutu B, Snow RW, Tshilolo, L, Zurovac D, Vanderelst D, and Speybroeck N (2018). True malaria prevalence in children under five: Bayesian estimation using data of malaria household surveys from three sub-Saharan countries. Malaria Journal, 17(1), pp.1-7.
- Rogan WJ, Gladen B, Estimating prevalence from the results of a screening test (1978). American Journal of Epidemiology, 107(1): 71–76, https://doi.org/10.1093/oxfordjournals.aje.a112510

4. SAMPLE SIZE CONSIDERATIONS

Careful consideration should be given to the sample size of a prevalence study to ensure the sample will produce estimates with adequate precision. The sample size required depends on features of the design (stratification and/or clustering), anticipated prevalence in the target population, the finite population size, and the desired level of precision. The R package “PracTools” is freely-available to the research community and includes functions for sample size calculations under multiple probability-based designs.

Design by AdrialDesigns.com

Because prevalence of SARS-CoV-2 infection tends to be fairly low in most populations, confidence intervals based on a normal approximation can undercover the true population prevalence. The Wilson method and log-odds methods are alternatives that have better coverage properties for prevalences close to the extremes of the parameter space. The "nWilson" and "nLogOdds" functions in the PracTools package calculate required sample sizes using these methods based on user-specified values of the targeted margin of error and population prevalence. These functions assume a simple random sampling design. Alternatively, the "power.prop.test" function within the Stats package in R can be used if the researcher is interested in powering the study based on a test of differences in prevalence between two non-overlapping groups when simple random sampling is used. The examples below demonstrate the use of the "nWilson" and "power.prop.test" functions for two different hypothetical designs.

Another reasonable method for power or sample size calculations, particularly for complex designs, is a simulation study. Researchers can simulate the finite population under a given set of assumptions, select a random sample from that finite population under the proposed sampling design, and calculate estimates based the sample. By repeating this process many times, the researchers can estimate power or precision empirically.

Regardless of the method used, it is important to remember that the sample sizes provided in the power or precision calculation reflect the number of participants in the study. When persons who are not eligible to participate are included on the sampling frame and/or nonresponse is anticipated, the sample size should be inflated to account for the anticipated eligibility and response rates for the study. For probability-based samples, it is good practice to select a larger sample than is anticipated being required and dividing the sample into random replicates that can be released over time if sampling targets are not achieved.

Design by AdrialDesigns.com

EXAMPLE CODE:

Example 1: Assume that a research team is conducting a seroprevalence study of the general population in a given county. They believe that population seroprevalence is approximately 8%, and would like to estimate seroprevalence from a probability-based sample with a corresponding 95% Wilson confidence interval with a margin of error of $\pm 2\%$. The sample will be recruited using address-based sampling, with n addresses selected via a simple random sample. As demonstrated in the R code below, the researchers need to obtain 713 participants to achieve their desired level of precision.

```
nWilson(moe.sw=1,alpha=0.05,pU=0.08,e=0.02)

$`n.sam`
[1] 712.077

$`CI lower limit`
[1] 0.06225363

$`CI upper limit`
[1] 0.1022536

$`length of CI`
[1] 0.04
```

Example 2: A research team is conducting a seroprevalence study among healthcare workers within the local healthcare system. They plan to stratify the sampling frame of healthcare workers by whether or not they work directly with COVID-19 patients and select random samples of n=500 within each stratum. They anticipate that seroprevalence among healthcare workers who do not work directly with COVID-19 patients will be 10%, and they want to know what the minimum detectable difference is assuming a Type I error rate of 0.05 and 80% power. Based on this design, the study would have 80% power to detect a significant difference if the true seroprevalence among workers who work directly with COVID-19 patients is 15.9%.

Design by AdrialDesigns.com

```
library(stats)

power.prop.test(n = 500, p1 = 0.10, p2 = NULL,

             sig.level = 0.05, power = 0.80,

             alternative = c("two.sided"))


    Two-sample comparison of proportions power calculation

              n = 500
             p1 = 0.1
             p2 = 0.1594911
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

REFERENCES/ADDITIONAL RESOURCES:

- Valliant R, Dever JA, and Kreuter F (2013). Practical tools for designing and weighting survey samples. New York: Springer.
- PracTools R Package: https://cran.r-project.org/web/packages/PracTools/index.html
- Stats R Package: https://www.rdocumentation.org/packages/stats/versions/3.6.2

Design by AdrialDesigns.com