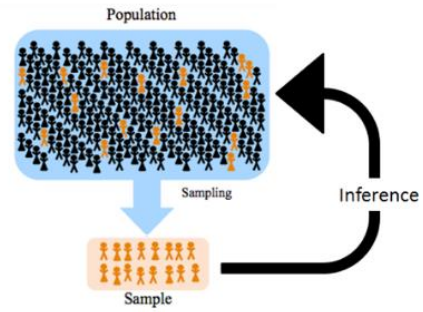
STATISTICAL APPLICATION IN SURVEY STUDY

Statistical analysis is widely used in Survey studies to provide timely information in a population of interest on important topics, such as demographic distribution, voter opinions, and disease prevalence. Analysis of survey study data requires some special attention because of complex sampling designs employed to obtain reliable and efficient population-level inference. Survey methodology allows scientific researchers to intentionally alter distributions of different subgroups in the survey sample by over- and/or under-sampling some subpopulations so that the survey sample groups are well represented to ensure reliable population-level estimates without resorting to extremely large samples.

SURVEY STUDY

THE ROLE OF STATISTICS IN SURVEY STUDY: Technologic advances in the internet and digital arena have greatly facilitated data collection for research and other purposes. In particular, surveys can be conducted to collected data instantaneously to provide important information for timely issues and topics. However, the paper to digital media transition has also created problems for data analysis. One particular issue is increased numbers of outliers, yielding uninterpretable and often biased results when analyzed using mean-based statistical models, including most popular models such as t-tests and regression. Rank-based methods such as the Mann-Whitney Wilcoxon ran sum test (MWWRST) and rank regression address this statistical problem without any subjective bias as in popular winsorized methods.

PREVIOUS STATISTICAL WORK IN SURVEY STUDY: As survey samples are not representative of populations of interest, standard statistical methods do not apply to such data. After Horvitz and Thompson's seminal work[1], many popular statistical methods have been extended to survey data and supported by popular statistical packages such as R, SAS, SPSS and STATA all provide support for such methods. For example, many popular SAS procedures have their survey counterparts to facilitate analysis of survey data, such as PROC SURVEYREG for linear and PROC SURVEYPHREG for Cox regression analysis. More recently, Lumley and Scott[2] developed an approach and an accompanying R package to extend the MWWRST to survey data for facilitating research. Their work is significant, as it represents the very first attempt to apply Horvitz and Thompson's inverse probability weighting technique to rank-based statistical models. However, their approach focused on testing the null of equal distribution. Although efficient for comparing two distributions, is very limited in practice, since the MWWRST is generally called for when the two-sample t-test is inappropriate due to outliers, in which case interest lies in comparing "centers" of two distributions, not equality of two distributions.



A NEW EXTENSION OF TRADITIONAL METHOD TO SURVEY DATA: By utilizing latest development in semi-parametric models, we developed an alternative MWWRST to compare mean ranks between two groups for survey data. Like mean and median, mean rank is a meaningful measure of the center of a distribution. Unlike its counterparts, mean rank for a group is calculated based on ranking observations from both groups and thus is not the same as median. Many erroneously interpret MWWRST as comparing two mean ranks as comparing medians of distributions. Although the two measures are identical for some special distributions, they are generally different. Thus, our work additionally clarified the correct interpretation of MWWRST and settled the debate on whether the MWW test really compares mediums of two distributions.

AN APPLICATION TO NHANES DATA: Anemia is a condition of decreasing in the total amount of red blood cells or hemoglobin in the blood. Researchers are interested in the association between serum copper and anemia[3]. We analyzed such associations based on the National Census and National Health and Nutrition Examination Survey (NHANES) data. The traditional two-sample t-test for survey data showed a significant association, but our approach did not, as it successfully addressed the artifacts of outliers. **Thus, researchers should be cautious about applying standard statistical models when the data include outliers.**

1. Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. J Am Stat *Assoc*. 1952;47(260):663-685.
2. Lumley T, Scott AJ. Two-sample rank tests under complex sampling. Biometrika. 2013;100(4):831-842.
3. Knovich MA et. al., The association between serum copper and anaemia in the adult second national health and nutrition examination survey (NHANES II) population. *Br J Nutr*. 2008;99(6):1226-1229.
Figure reference: https://online.stat.psu.edu/stat200/book/export/html/21