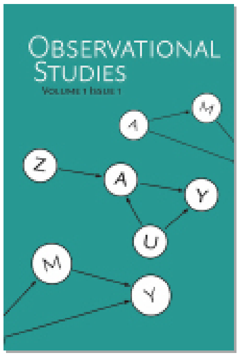
Comment on Cochran's "Observational Studies"

Joseph L. Gastwirth, Barry I. Graubard

Observational Studies, Volume 1, Number 1, 2015, pp. 173-181 (Article)

Published by University of Pennsylvania Press
DOI: *https://doi.org/10.1353/obs.2015.0015*



Observational Studies 1 (2015) 173-181 Submitted 4/15; Published 8/15

Comment on Cochran's "Observational Studies"

Joseph L. Gastwirth **jlgast@gwu.edu**
Department of Statistics, George Washington University
Washington, DC 20052, USA

Barry I. Graubard **graubarb@exchange.nih.gov**
Division of Cancer, Epidemiology & Genetics, National Cancer Institute
Bethesda, Maryland 20892, USA

1. Introduction

The timelessness of Prof. Cochran's contributions to the planning, design and analysis of both large scale surveys and observational studies is exemplified by his 1972 paper summarizing many years of research. Prof. Small should be thanked by the statistics community for devoting a special section of the new journal *Observational Studies* to bring this important work to the attention of the next generation of statisticians and data scientists. Researchers in almost every field will benefit from reading the advice given in the paper at the very start of thinking about a study, whether randomized or observational. Our commentary will focus on differences between studies designed to make inferences applicable for the general population or studies carried out to understand what occurred in the population studied, i.e. the study population is the population for which inferences will be made. In the first and most common setting, the importance of Cochran's observation that one needs to consider whether the study population differs in some important ways from the general population will be illustrated by reviewing studies concerning the relationship between obesity and more generally body weight and mortality and morbidity. While that issue does not arise in the second setting, frequently occurring in legal cases dealing with discrimination or violation of occupational safety and health rules, the wisdom of Prof. Cochran's recommendations on analytic techniques and methods for controlling for the potential effect of confounders are very relevant to the proper analysis of statistical evidence.

2. Inference from samples from a population

Cochran points out that the target population should be identified and that a probability sample of the target population be collected. However, he points out that a probability sample may not be possible, e.g., because of cost and operational considerations, so that many studies obtain a sample from a population that differs somewhat from the target population. For example to study the association of body weight (actually body weight adjusted for height or body mass index (BMI), i.e., weight in kg divided by the square of height in meters) with all-cause mortality, researchers have used existing cohorts of adults such as the National Institutes of Health (NIH)-AARP Diet and Health Study (Adams et al., 2006), which we will call the NIH-AARP Study. The NIH-AARP Study sent a questionnaire in 1995-96 to all members of the AARP 50-71 years old who resided in six U.S. states (California, Florida, Louisiana, New Jersey, North Carolina, and Pennsylvania) and two metropolitan areas (Atlanta and Detroit). The questionnaire collected self-reported body weight, height along with other relevant covariates such as smoking, alcohol consumption and physical activity that were incorporated in the analysis to remove the potential for confounding in the estimated association of BMI and mortality. The sample studied had information on the 18% of male and female members of AARP (n= 567,169) who returned the questionnaires. The low response rate also raises important statistical concerns about the generalizability of the conclusions to the population of AARP members, much less the entire adult U.S. population 50-71 years old.

The Adams et al. article states "Even against the background of advances in the management of obesity-related chronic diseases in the past few decades, our findings suggest that adiposity, including overweight, is associated with an increased risk of death" and compares their results to those reported by Flegal et al (2005). The earlier study used national probability samples from the National Health and Nutrition Examination Survey (NHANES) to construct US representative cohorts and did not find a an increased risk of death in overweight ($25 \leq \text{BMI} < 30$) individuals. The NIH-AARP Study does not state restrictions on the population to which their conclusions are applicable and implies that they are valid for the entire US population. Investigators (Durazo-Arvizo et al 1997; Calle et al 1999) have reported that the BMI mortality relationship may vary by race, so those groups should be appropriately represented in the study sample. Although membership in the AARP is open to the entire population, it is not a representative sub-population of the over 50 population of the nation as members must pay annual dues. Further, the low response rate potentially exacerbates the non-representativeness of the NIH-AARP sample.

Cochran's recommendation that discussing how differences between a study sample and the target population, such as the US population, may affect the interpretation of the inferences drawn from it is very important. The BMI and mortality relationship found in the NIH-AARP Study may not be generalizable to the entire US population as its race/ethnic mix differs from that of the entire US. The 1995 US census projections for race/ethnicity distribution of 50-69 year olds (the closest age range to the NIH-AARP available) were 80.7%, 9.5%, 6.5%, 3.3% (see http://www.census.gov/prod/1/pop/p25-1130/p251130.pdf) compared to the NIH-AARP race/ethnicity distribution of 91.6%, 3.7%, 1.8%, 1.6% for white, African-American, Hispanic and Asian, Pacific Islander or Native American, respectively. Not surprisingly, whites are over-represented in the NIH-ARRP study; notice that all three minority groups are under-represented by a factor of two. The use of volunteers, i.e., nonrandom samples, from selected populations is typical of many epidemiologic studies; examples are plentiful, e.g., the Harvard University Nurses Health Study `http://www.channing.harvard.edu/nhs/`, the National Cancer Institute U.S. Radiologic Technologists Cohort
`http://dceg.cancer.gov/research/who-we-study/cohorts/us-radiologic-technologists`, and the City of Hope National Medical Center California Teacher Study
`https://www.calteachersstudy.org/study_data.html`. Even though the sample size of these cohorts is quite large, the potential bias resulting from having specialized samples of a target population (e.g., the US population) is not ameliorated by the small standard errors of estimates of association obtained from large samples. This important point is

often ignored by epidemiologists even though statisticians are well aware of the potential magnitude of this type of bias, e.g., from the Literary Digest pre-election poll in 1936. For conclusions concerning a relative risk obtained from a sampled population differing from the target one, the risk model fit to the data needs to be correctly specified with the correct functional form of the relationship of the response to the covariates along with appropriate interactions and information on all major covariates should be obtained. Only then will there be a good chance of estimating the risk accurately for a target population. Furthermore, the sample should span the same distribution of covariates as the target population, e.g., if the relative risks differ by age group the age groups of the target population should be represented in the sample population.

From a public health policy point of view, the results from cohort studies or other types of epidemiologic studies are most useful if they are generalizable to the target population of interest. In the Adams et al. paper and many other epidemiologic analyses, population attributable risks (PAR) for an "exposure" are used to estimate how many cases or deaths would be prevented if the exposure was eliminated or reduced by a suitable intervention. If the estimated relative risks from a particular study are not applicable to the target population then the estimated PARs could be misleading resulting in a misallocation of resources that may be directed to more important public health exposures, e.g., preventing smoking, or warning the public of a risk of a serious disease, e.g. Reye syndrome, from a frequently used product.

3. Inference for the study population

Although the objective of most statistical surveys and studies is to draw inferences from a subset, ideally a random sample, of a population that will apply to the much larger population, in some important applications one is concerned with drawing conclusions that will apply only to the study population. In many legal cases the question addressed by the statistical analysis can shed light on concerns of the appropriateness of the practices of a specific employer or firm. For example, in a fair trade case the question may focus on whether a particular exporter "dumped" or sold goods below cost, which is unfair to the importing nation's producers; in an equal pay case the issue is whether female employees are paid the same as similarly qualified males. In both situations, the conclusions will only apply to the particular firm or employer, i.e. if it turns out that the firm did not dump goods or that the employer underpaid female employees by $2.00 an hour, those conclusions will not be considered in a similar case involving a different firm nor would they imply that females employed in similar jobs throughout the nation are under-paid by an average of $2.00 an hour. This section will show that several of Prof. Cochrans wise suggestions and guidelines are very useful in analyzing this type of observational study but also have been misinterpreted by "experts" and courts as they were developed for situations where the ultimate inference will apply to a much larger population than the one studied.

In the legal setting where one has data for the entire finite population for the period under review, summary statistics calculated from the data are in fact population quantities. Statisticians often impose a probability model to aid in interpreting and understanding the evidentiary strength of a difference in averages or percentages. For example, in an equal employment case concerning the fairness of an employers promotions, suppose that 2 of

15 (13.3%) eligible female employees and 12 of 23 (52.2%) eligible males were promoted during the relevant time period. There is no sampling error involved; however, as an aide to interpreting the data statisticians often assume that the promotions are *randomly* chosen from the pool of eligible employees. Assuming that all other relevant factors, e.g. seniority are balanced in both groups, the number of females among the 14 promotions follows a hypergeometric distribution and Fisher's exact test yields a p-value of .02 (two-sided). Notice that the proportion of women promoted is about one-fourth the corresponding proportion of male employees, which is clearly meaningful. The statistical test informs the court that the data is unlikely to occur if promotions were randomly selected from the eligible pool. Thus, we infer that the gender of an eligible employee affected their chance of promotion. Courts then require the employer to justify their promotion process.

Notice that the total sample size in the above situation is *much* smaller than in the applications discussed by Prof. Cochran. Unfortunately, courts often ignore data sets referring to the complete set of eligible employees and simply say the sample is too small. For example, the decision in the age discrimination case, *Fallis v. Kerr-McGee Corp.* 944 F.2d 743 (10th Cir. 1991), stated that the "sample" of 51 employees was too small.[1] A related problem is that expert witnesses have convinced courts that samples of 200-400 may be needed to subject data comparing the pass rates of minority and majority pass rates on a pre-employment exam to assess whether it has a disparate impact on the minority applicants.[2]

Because courts have not encouraged the analysis of data pertaining to a seemingly small population, they may not fully appreciate the meaning of a simple statistical summary. The case *Chappel v. Ayala*[3] currently being considered by the U.S. Supreme court provides an illuminating example. The case concerns whether the lower courts properly considered a defendants Batson allegation that the prosecutor discriminated against minorities by removing all seven minority members from the venire of potential jurors through peremptory challenges. Although the main legal issues concern the propriety of the trial judge excluding the defendant's lawyer from part of the proceedings where the prosecutor explained why the minorities were challenged and the apparent loss of many questionnaires potential jurors filled out, the courts might have benefitted from a formal statistical analysis of the data. Even Judge Callahan who dissented from the 9th circuits opinion granting the defendant a new trial noted "The only indicia of possible racial bias was the fact seven of the eighteen peremptory challenges exercised by the prosecutor excused African-American and Hispanic jurors." To properly interpret this information one needs to know the number of non-minorities who were on the venire. The majority opinion noted that in the case, each side could remove 20 members of the venire by peremptory challenge when the jury of 12 was chosen and then had six more peremptory challenges to use when the six alternates were chosen. Thus, there must have been at least seventy individuals on the venire in order for the court to end with a jury of twelve and six alternates. To *maximize* the proportion of

[1] In the case, 3 of 9 employees over 40 were fired in contrast to 4 of 42 employees under 40. Analyzing the data with Fisher's test yields a non-significant result (one-sided p-value = .095), which would support the courts decision and avoid making an "ad hoc" judgement that a sample of 51 is too small.

[2] See Lopez v. Lawrence (D. Mass.) 2014 U.S. Dist. LEXIS 124139.

[3] The Supreme Court granted certiorari in Chappell v. Ayala, 2014 U.S. LEXIS 7094 (U.S., Oct. 20, 2014) and will review the decision Ayala v. Wong 756 F.3d 656 (9th Cir. 2014); 2014 U.S. App. LEXIS 3699.

minorities in the pool from which the jury of twelve were chosen, let us assume that the trial court proceeded in two stages: first, selecting the jury and then the alternates. Allowing for each side to have twenty peremptory challenges, the minimum size of the panel from which the twelve jurors were chosen is 52, of whom 7 were minority. The prosecution actually removed 18 members of the panel, all seven minorities and eleven whites. Applying Fishers exact test shows that the probability that a random sample of 18 taken from a pool of 7 minorities and 45 whites would include all 7 minorities is .00024 or about 1 in 4000.[4] This is quite a significant result, which suggests that the court should carefully examine the reasons the prosecution offers to justify its challenges when the judge compares the characteristics of the minority members excluded with the majority members who were not excluded to see whether the offered reasons were applied to all members of the venire.[5]

Prof. Cochran emphasized the usefulness of matching and stratification methods and they are especially appropriate when the results of one's analysis needs to be explained to a non-statistician as they can understand that the factors used in the matching/stratification process are controlled for. In other words proper matching and stratification can simplify analysis to examining possibly simple means or proportions when otherwise, for instance, a less intuitive regression method may be used to conduct analyses that adjust for the stratification or matching variables.

If the concomitant factor used in the matching process is ordinal, however, one may lose some relevant information. As an example consider the pay data in Table 1 on the following page from EEOC v. Shelby County Government, a case concerning whether women clerical employees in the county's criminal court were discriminated against in pay and promotion. In the opinion, the judge noted that judges are very familiar with the duties of these clerical workers and found unequal pay after considering the pay data in Table 1, stratified into four seniority levels. Although the data is so clear a formal statistical test was not required, Gastwirth (1992) applied the Mann-Whitney form of the Wilcoxon test to the data in each strata and combined the results using the van Elteren procedure. The result was highly significant (p¡.001). One feature of the data, however, is ignored in this analysis. Notice that some men are paid more than women with noticeably more seniority. For example, D.V., a male is paid more than the four females who have higher seniority (i.e., F.R., T.D., P.B., and P.E.) and B.W., another male who has even less seniority than D.V., is paid more than three of those females and the same as the fourth. This phenomenon held true even in 1988, five years after the charge was filed (Gastwirth, 1992).

[4] If the trial court started with a panel large enough for it to select twelve jurors and six alternates, then the minimum size would be 70 and the probability that all seven minorities would be in a random sample of 18 from this larger pool would be 2.55×10^{-5} or just over one in 40,000. Unfortunately, none of the opinions reports the full data set or provides a detailed description of the original jury selection procedure.

[5] In United States v. Omoruyi, 7 F.3d 880 (9th Cir. 1993), the prosecutor peremptorily challenged the two single minority females and the defendant raised a Batson claim after the second one was removed. The trial judge accepted the prosecutor's claim that he removed them because they were single. The appellate court noted that the prosecutor had not peremptorily challenged single, unmarried men in the jury panel and granted the defendant a new trial. In contrast, in Alviero v. Sam's Warehouse Club Inc. 253 F.3d 933, 940-41 (7th Cir. 2001) the court accepted the prosecutor's explanation of the removal of all three female members of the jury panel on the basis of their limited work experience and level of education even though some males with similar educational backgrounds but more work experience but more education were not challenged.

Table 1. Pay Data for Male and Female Employees Clerical Employees of Shelby County Criminal Court, and Estimated Damages for Female Employees using the Peters-Belson Approach

Initials of Employee	Gender	Hire date	Salary in 1983 (dollars per month)	Estimated damages for female employees for 1983 (dollars per month)
F.R.	F	5/73	1474	203.61
J.P.V.	M	9/73	1666	
T.D.	F	1/74	1403	227.50
C.H.	M	1/74	1666	
P.B.	F	5/74	1403	203.94
L.A.	M	5/74	1548	
C.C.	M	5/74	1548	
P.E.	F	8/74	1403	186.27
G.V.	M	9/74	1548	
T.P.	F	5/75	1112	424.27
G.L.	F	1/76	1306	183.15
S.B.	F	2/76	1336	147.26
D.V.	M	3/76	1548	
J.B.	F	9/76	1336	106.04
B.W.	M	1/78	1474	
B.D.	F	9/78	1000	300.69
B.P.	F	10/79	1000	224.12
J.A.	M	10/79	1157	
F.D.	M	8/82	1000	
P.S.	F	9/82	929	88.99
M.D.	F	12/82	929	71.32
V.H.	F	1/83	929	65.43
S.C.	F	7/83	800	159.10

The Peters-Belson (Peters, 1941; Belson, 1956) approach to regression, discussed in Cochran and Rubin (1973), was used by Gastwirth and Greenhouse (1994) to analyze this data. First, one fits a model relating the salaries of male employees to their seniority level. Then one predicts the salary a female employee would receive had they been paid according the male equation. The differences D_i for each female estimate the shortfall (if negative) in their salary and $Z = \bar{D}/\sqrt{V(\bar{D})}$ where $\bar{D}$ is the average of the D_i and $V(\bar{D})$ is the variance of $\bar{D}$, and Z is approximately normally distributed in large samples and a t-distribution in small ones. For the Shelby data the model was a linear regression predicting a worker's salary in 1983 from the number of months they had worked. Table 1 displays the salary and date hired and gender data that we use here; see Table 7 in Gastwirth (1992) for this data and salary data for other years. The observed average shortfall $\bar{D} = \$185.12$ in the monthly salary of females has a standard error of 35.62 resulting in two-sided p-value $< .001$. Another analytic approach that does not require the assumption that the errors in the regression model follow a normal distribution and logically follows from the idea that one is imposing a probability model on the data is to apply a permutation test. A complete permutation test would consist of swapping the gender labels and repeatedly applying the Peters-Belson approach to each relabeled data set. As there are 9 males and 14 females, there are $\binom{24}{9} = 1,307,504$ ways to relabel gender in the Shelby data. For computational purposes we randomly selected (without replacement) 1,000 relabeled data sets and found only 3 of the $\bar{D}$ across the 1,000 to be as large or larger in absolute value as the observed shortfall, yielding a two-sided p-value of 0.003. Table 1 shows the Peters-Belson estimate of D_i for each female employee is negative, which illustrates the unfairness of the pay system examined and provides an estimate of the amount of money each woman deserves. Other uses of permutation methods in least squares are described in Sprent (1998).

The Peters-Belson (PB) method is related to the use of counterfactuals in the Neyman-Rubin (Holland, 1986; Morgan and Winship, 2007) approach to causal inference if one considers the PB estimated salary for each female obtained from the male equation as the estimated salary of her "male counterfactual." This predicted salary, however, may not be the salary of any of the male employees; rather it is a "statistical match" in the terminology of Peters (1944). The accuracy of the shortfall obtained from PB regression depends on the appropriateness of the model and the completeness of the information on the covariates. In the context of an "Equal Pay" case, the employer knows the relevant factors used and in determining salaries and the relative weight given to each of them and should ensure that accurate information on them is obtained and retained.

4. Summary and future thoughts

Very few publications remain highly relevant to their subject after forty years have passed. Professor Cochran's 1972 paper and his earlier work, which is summarized in it, are in that special category. Every investigator should review the recommendations on the need to have a clear statement of the objectives of a study when planning and designing a study and follow his suggestions, e.g. have a pilot study, at those stages. His discussion of the various methods for removing the effect of confounders remains the basis of much current research (Rosenbaum, 2002).

Our comments focused on the value of Cochran's emphasis on the need to consider the effect of differences between the study population and the population to which the inferences drawn from the study will be applied and the situation when the study group is the entire population. In the context of examining the complete population, especially in a legal case, Prof. Cochran's concern with stability of the relationship, presumably over time, is less important than in the usual setting where one desires to draw inferences valid for a much larger population from a sample and learn about the underlying mechanisms producing the response. In an equal employment case, one's focus is on what happened during the few years in which the employer used the practice (job requirement, pay decision process) under review. Indeed, quite often an employer will change policies in response to a claim so that the earlier relationship between salary and gender or race and other covariates may well change.

In practice, almost no large scale study will be "perfect," the statistical model will generally be a good "approximation" of the relationship of the response to the predictors and there will be errors of measurement and a potentially relevant covariate may be omitted. Readers should be aware of the usefulness of sensitivity analysis (Rosenbaum, 2002) and in particular, the importance of the Cornfield inequality (Cornfield et al., 1959, Gastwirth and Greenhouse, 1995) in assessing whether a possible omitted factor can explain a statistically significant difference between two groups. Briefly, Cornfield gave conditions on the strength and imbalance or differences in the prevalence the omitted variable must satisfy in order to explain a relative risk. The result has been used by Gastwirth (1992) to show that judges who required the party suggesting that an observed difference or relative risk was due to an omitted variable to submit an analysis including that factor were correct.

In view of recent interest in the issue of reproducibility of scientific studies, basing inferences that will be applied to the target population on random samples from it has the advantage that investigators using different random samples should arrive at similar results, within sampling variation.

References

Adams K.F., Schatzkin A., Harris T.B., Kipnis V., Mouw T., Ballard-Barbash R., Hollenbeck A. and Leitzmann M.F. (2006). Overweight, obesity, and mortality in a large prospective cohort of persons 50 to 71 years old. *New England Journal of Medicine*, 355, 763–78.

Belson W.A. (1956). A technique for studying the effects of a television broadcast. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 5, 195-202.

Calle, E.E., Thun, M.J., Petrelli, J.M., Rodriguez, C. and Heath, C.W. Jr. (1999). Body-mass index and mortality in a prospective cohort of U.S. adults. *New England Journal of Medicine*, 341, 1097-1105.

Cochran, W.G. and Rubin, D.B. (1973). Controlling bias in observational studies, a review. *Sankyha* (A), 35, 417–446.

Cornfield, J.C., Haenzel, W., Hammond, E.C. , Liliefield, A.M., Shminkin, M.B. and Wynder, E.L. (1959). Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer Institute*, 22, 173–203.

Durazo-Arvizu R., Cooper R.S., Luke A., Prewitt T.E., Liao Y. and McGee D.L. (1997). Relative weight and mortality in U.S. blacks and whites: findings from representative national population samples. *Annals of Epidemiology*, 7, 383–395.

Flegal K.M., Graubard B.I., Williamson D.F. and Gail M.H. (2005). Excess deaths associated with underweight, overweight, and obesity. *Journal of the American Medical Association*, 293, 1861–1867.

Gastwirth, J.L. (1992). Methods for assessing the sensitivity of statistical comparisons used in title VII cases to omitted variables. *Jurimetrics*, 33, 19–34.

Gastwirth, J.L. (1992). Statistical reasoning in the legal wetting. *American Statistician*, 46, 55–69.

Gastwirth, J.L. and Greenhouse, S.W. (1995). Biostatistical concepts and methods in the legal setting. *Statistics in Medicine*, 14: 1641–1653.

Holland P.W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81, 945–970.

Morgan, S.L. and Winship, C. (2007). *Counterfactuals and Causal Inference: Methods and Principles for Social Research.* Cambridge University Press, Cambridge.

Peters C.C. (1941), A method of matching groups for experiment with no loss of population. *The Journal of Educational Research*, 34, 606-612.

Sprent, P. (1998). *Data Driven Methods in Statistics.* Chapman & Hall, London.