Qualifying Exam

Casey Breen

2020-12-31

Contents

Chapter 1

Summary

This set of memos covers two reading lists:

- Causal Inference and Population Studies
- Social Networks

The general format of the memos will be as follows:

- Key Background
- Data / Methods
- Research Question
- Argument / contribution
- Open Questions

Chapter 2

Causal Inference and Population Studies

2.1 Foundations

Adler, Nancy E., and David H. Rehkopf. 2008

Citation

Adler, Nancy E., and David H. Rehkopf. 2008. "U.S. Disparities in Health: Descriptions, Causes, and Mechanisms." Annual Review of Public Health.

Key Background

- Eliminating health disparities in a fundamental goal of public health research and practice.
- Studying health disparities is challenging due the (i) definition of a disparity and (ii) ability to attribute cause from association.

Methods

- Extensive literature review of both descriptive + causal methods for identifying substantice applications and Annual Review Article — summarizes key studies related to health disparities.

Research Question

- What are the key research has been conducted on the causes and mechanisms of health disparities in the US, and how is this research conducted?
- What are the causes of health disparities in the US and why what are the specific pathways and mechanisms driving these disparities?

Argument / contribution

- A health disparity is a broad term loosely defined. The broadest definition refers to health differences that occur with respect to gender, race or ethnicity, education, income, geographic location, or sexual orientation. Health disparities result from both biological differences and social disparities, but the latter has a greater effect and in avoidable.
- Data limitations often preclude the study of SES and health—particularly the intersection of SES and Race/Ethnicity. Further, SES indicators may have different meanings for different groups (e.g., Blacks and Hispanics have lower wealth than non-Hispanic Whites and Asians at a given income level), further complicating "controlling" for a covariate.
- Descriptive understandings are important for (i) understanding short and long-term trends in mortality disparities, (ii) sparking causal investigations of health disparities, (iii) allocating resources to reduce disparities in specific diseases, and (iv) increasing public awareness.
- Analytic approaches for establishing causality include propensity score matching, instrumental variables, time-series analysis, causal structural equation modeling, and marginal structural models.

- Identifying the specific pathways and mechanisms by which SES and race/ethnicity affect health can strengthen causal claims and help target health interventions. For example, differential exposure to stress, particularly repeated exposure, is one of multiple pathways identified in the literature.

Unanswered Questions

- How do we collect data with adequate measures of SES, demographic covariates, and health outcomes to identify understand the mechanisms and pathways?

Xie, Yu. 2013.

Citation

Xie, Y. 2013. "Population Heterogeneity and Causal Inference." Proceedings of the National Academy of Sciences 110(16):6262–68. doi: 10.1073/pnas.1303102110.

Key Background

- The core objective of social science research is to understand the (ubiquitous) population heterogeneity. (Not identify abstract and universal laws).
- Causal inference with observational data is only possible with strong assumptions.

Methods

- Conceptualize selection into treatment as a dynamic process.
- Composition bias is generated by a dynamic process when the treatment proportion changes. Demonstrated via simulation.

Research Question

- What is the relationship between population heterogeneity and causal inference? How can composition bias arise from population heterogeneity?

Argument / contribution

- It is impossible to draw causal inference at the individual level due to heterogeneity.
- Composition bias, a form of selection bias, arises often in the real world. For example, if the administration of a medical treatment or social intervention is done on a graduated schedule, where participation is need based. Individuals selected ata later states will have lower ATEs because they are coming from a less responsive sub-population.

Key Findings

- Composition bias occurs because units with a higher intrinsic propensity of treatment are more likely to be over-represented when the treatment proportion is small.
- Researchers should be mindful of the specific subgroups of interest when deriving and interpreting average causal estimates from potentially heterogeneous subgroups.

Unanswered Questions

- When can we establish external validity for research results in?

Gangl, Markus. 2010

Citation

Gangl, Markus. 2010. "Causal Inference in Sociological Research." Annual Review of Sociology 36(1):21–47. doi: 10.1146/annurev.soc.012809.102702.

Key Background

- The counterfactual model provides a natural framework for clarifying the requirements for valid causal inference

Methods

- Summarizes recent literature on causal inference literature (estimands, quasi-experimental designs, diff-in-diff, instrumental variable estimation, semi and non-parametric methods)

Research Question

- What's the role of causal inference in sociological research?

Argument / contribution

- Traditional setup in sociology papers is insufficient—comparing regression specifications and testing competing hypotheses is unlikely to identify any causal effect of interest.
- While the benefits of RCTs are widely understood, regression using observational data is the workhorse of sociology research. Lit review provides a sobering view of the efficacy of drawing causal conclusions from regression analysis, as practiced in sociology.
- Counterfactual framework doesn't only apply to explicitly manipulable treatments—it can also apply to non-manipulable factors such as gender, race, or class.
- Standard treatment of effect estimates are local — historically and situationally contingent.

Key Findings

- The availability of longitudinal data and informative natural experiments will allow for causal identification within an area plagued by confounders.

Petersen, Maya L., Sandra E. Sinisi, and Mark J. van der Laan. 2006.

Citation

Petersen, Maya L., Sandra E. Sinisi, and Mark J. van der Laan. 2006. "Estimation of Direct Causal Effects." Epidemiology 17(3):276–84.

Key Background

- Most common problems in epidemiology (and social science more broadly) involve estimating the effect of an exposure on an outcome while blocking the exposure's effect on an intermediate variable. Estimation of direct effects is typically the goal of research attempting to estimate the causal pathways for which a treatment causes an outcome.
- Controlled direct Effect: hold intermediate variable at a fixed level. In contrast, a natural direct effect would measure the effect of the exposure, blocking the exposure's effect on intermediate background but allowing the intermediate to vary among individuals.

Methods

Research Question

What's the difference between a controlled and a natural direct effect and how can the natural direct effect be estimated (and with which underlying assumptions)?

Argument / contribution

- Natural direct effect of an exposure on an individual is defined as the difference in counterfactual outcome if an individual was unexposed vs. exposed allowing the intermediate to remain at counterfactual level under no exposure. For a controlled direct effect, the intermediate would be set to a fixed value for all persons.
- To estimate the direct effect in the whole population, one can take the weighted average of subgroup-specific direct effects with the weight fort a given subgroup determined by the relative size of the subgroup in comparison with the population.
- Formal assumptions for estimating controlled and natural effects

$$A \perp Y_{AZ}|WZ \perp Y_{AZ}|A, W \tag{2.1}$$

In words, this means no unmeasured confounders of either the effect of the exposure on the outcome of the effect of the intermediate on the outcome.

Additionally, to identify natural direct effects, we need:

$$A \perp Z_a|W \tag{2.2}$$

which in words says there are no unmeasured confounders of the effect of the intermediate on the treatment.

Second, within subgroups defined by covariates, the level of the intermediate variable in the absense of exposure does not tell us anything about the expected magnitude of exposure's effect at a controlled level of the intermediate variable. (Direct effect assumption)

$$E(Y_{az} - Y_{0z}|Z_0 = z, W) = E(Y_{az} - Y_{0z}|W) \tag{2.3}$$

Key Findings

- The barriers to estimation of direct effects are not as great as have been previously suggested. Researchers are encouraged to estimate direct effects which giving appropriate consideration to relevant assumptions and interpretations of their estimates.

Imai, Kosuke, Luke Keele, Dustin Tingley, and Teppei Yamamoto. 2011.

Citation "Unpacking the Black Box of Causality: Learning about Causal Mechanisms from Experimental and Observational Studies." American Political Science Review 105(4):765–89. doi: 10.1017/S0003055411000414.

Key Background

- Many empirical studies focus on establishing whether on variable affects another and fails to explain how such a causal relationship arises. This "black box" approach to causality has been criticized across disciplines.
- A causal mechanism is a process in which a causal variable of interest influences an outcome.
- Quantitative investigation of causal mechanisms is based on the estimation of causal mediation effects.

Methods

- Review of causal inference literature, particularly as it applies to the estimation of causal mediation effects.

Research Question
- What methods and assumptions can researchers use to identify the causal mechanisms behind observed causal effects?

Argument / contribution

- Consistency assumption—potential outcomes must take the same values as long as the treatment and mediator values are the same—is especially important in the analysis of causal mechanisms. Even if experimental designs involve the manipulation of mediators, one must assume that subjects would respond in the same way if those values of mediators were spontaneously chosen by the subjects themselves.
- Instrumental variables: assume that there is no direct effect on the outcome and effects all units in one direction. However, this often leads to the "black box" approach to causal inference, where insufficient attenuation is paid to causal mechanisms.

- Given the difficult of studying causal mechanisms, some researchers believe that a focus should be place on identifiation of causal effects and not causal mechanisms.

Key Finding / Conclusion

- Paper demonstrates three ways to move forward in research into causal mechanisms. First, the potential outcomes model of causal inference. Second, a sensitivity analysis to evaluate robustness of conclusions. Finally, new research designs for experimental and observational studies can reduce need for untestable conclusions. These new methods allow the black box of causality to be unpacked, going beyond the estimation of causal effects.

Rothman, Kenneth J., and Sander Greenland. 2005.

Citation Rothman, Kenneth J., and Sander Greenland. 2005. "Causation and Causal Inference in Epidemiology." American Journal of Public Health 95(S1):S144–50. doi: 10.2105/AJPH.2004.059204.

Key Background

- Concepts of cause and causal inference are largely self-taught from early learning experiences.
- Causal inference in epidemiology is more an exercise in measurement of an effect rather than a criterion-guided process for deciding whether an effect is present of not.

Methods

- Literature review

Research Question

- How should we think about causation and causal inference in epidemiology?

Argument / contribution

- Analogy: turning a light switch "on" will lead to light going on. However, the complete causal mechanism is much more intricate (e.g., bulb, wiring, electricity). Such basic models cannot serve as the basis for scientific theory, and a better starting place is a general conceptual model.
- A given disease, for example, can be caused by more than one causal mechanism (multicausality). Thus, falling on an icy oath leading to a broken hip may be just part of a complicated causal mechanism that involves many component causes. Takeaway: most identified causes are neither necessary nor sufficient to produce disease.
- All scientific work are only tentative formulations of a description of nature, even when work is carried out without error.

Key Findings

- While all studies will have error, the key is to quantify the errors. There are no qbsolutet criteria for assessing the validity of scientific evidence, it is will possible to assess the validity of a study, although this requires familiarity with the study and subject matter.

Unanswered Questions

- What here isn't obvious? (I'm a bit confused by this paper.)

Robins, J. M., M. A. Hernán, and B. Brumback. 2000.

Citation

Robins, J. M., M. A. Hernán, and B. Brumback. 2000. "Marginal Structural Models and Causal Inference in Epidemiology." Epidemiology (Cambridge, Mass.) 11(5):550–60. doi: 10.1097/00001648-200009000-00011.

Key Background

- In observational studies, standard approaches are biased when there exists time-dependent confounders also affected by previous treatment.

Methods

Theory and worked example

Research Question

- How do we use MSMs to estimate the causal effect of a time-varying exposure or treatment on a dichotomous outcome? Specific example: what is the cumulative effect of zidovudine (AZT) treatment on the probability of having undetectable high RNA levels. modifited by baseline modifier of interest.

Argument / contribution

- IPTW relies on the consistent estimation of treatment mechanisms. If treatment mechanism g_0 is consistently estiamted, the resulting IPTW point estimates will be inconsistent, leading to biased inference.
- IPTW is easy to implement, and may be easier to estimate the probability of a binary exposure as the function of the past than estimate the Q compoment of the likelihood.
- IPTW estimator is no efficient, and is very susceptible to violations of the positivity assumptions.

Unanswered Questions

- When do other estimators outperform the IPTW estimator? Specifically, which estimator does well in the presence of near positivity violations.

2.2 Applications

Brand, Jennie E., and Yu Xie. 2010.

Citation

Brand, Jennie E., and Yu Xie. 2010. "Who Benefits Most from College?: Evidence for Negative Selection in Heterogeneous Economic Returns to Higher Education." American Sociological Review 75(2):273–302. doi: 10.1177/0003122410363567.

Key Background

- Positive selection is assumed for college education: those who are most likely to select into college benefit the most from college. This results from a utility maximization paradigm for economic factors (e.g., only attend college if economic returns outweigh the costs.)
- This paper introduces an alternative theory, the negative selection hypothesis: those who are most least likely to attend college are the most likely to benefit. Given the family and individual attributes are associated with the attainment of higher education, mechanisms influencing college attainment may differ by social background. For example, there is much heterogeneity in the financial burden and the family pressure of deviating from cultural norms based on one's socioeconomic background.
- There are two types of selection bias in observational data. First, heterogeneity in preexisting conditions or attributes associated with both the treatment condition and the outcome. Second, heterogeneity in treatment effects, i.e. systematic differences between those who do and do not attain a college education in the causal effect of a college education on earnings.

Research Question

- How can we adjudicate between the positive and negative selection hypothesis using causal propensity-score matching methods?

Methods

- Three step approach: (1) Invoke randomization assumption and assume no additional confounders between persons who do and do not complete college (2) Summarize differences in covariates between college and non-college goes using estimated propensity score matching. (3) Revisit the ignorability assumption using auxiliary analyses that aid the interpretation of results

Argument / contribution

- Ignorability assumption: potential outcomes are uncorrelated with treatment status, conditional on observed covariates. Can never be verified and should not be taken as true in practice for observational data.
- As only one outcome is actually observed, causal inference is impossible at the individual level and always requires analysis at the group level on the basis of some homogeneity assumption.
- We can't condition on all covariates **X, due to the curse of dimensionality; however, we can condition on a propensity score estimator.**
- **Comparing slopes of propensity score stratum and treatment effect of college completion on earnings of the life course shows support for negative selection hypothesis at each observed stage of the life course.**
- **When models were restricted to a more limited set of covariates, there was evidence for positive selection.**

Key Findings

- Individuals from the most disadvantaged social origins and the lowest ability and achievement, actually completed college, they benefited the most in terms of labor market outcomes.
- Heterogeneity in response to a treatment in a norm rather than an exception

Unanswered Questions

- How does the amount, quality, major, etc. of education influence labor market outcomes?
- More selectivity for those *less* likely to attend college?

De Neve, J. E., and A. J. Oswald. 2012.

Citation

De Neve, J. E., and A. J. Oswald. 2012. "Estimating the Influence of Life Satisfaction and Positive Affect on Later Income Using Sibling Fixed Effects." Proceedings of the National Academy of Sciences.

Key Background

- The causal effect of income on psychological well-being has been well studied across the social sciences, but reverse causality is arguably understudied.
- Positive correlation between "cheerfulness" measured in a sample of elite college students and their income levels 19 y later.

Research Question

- What's the relationship between income and happiness? What is the causal direction and what are the mediating pathways between income and happiness.

Methods

- Individual fixed effects and sibling fixed effects regression models controlling for self-esteem and life satisfaction using data from Add health.

- Sobel-Goodmand method for mediation analysis: (i) X significantly predicts M, (ii) X significantly predicts Y in the absence of M, (iii) M significantly predicts Y controlling for X, and (iv) the effect of X on Y shrinks upon addition of M.

Argument / contribution

- The most significant mediating pathways include obtaining a college degree getting hired and promoted, higher degrees of optimism and extroversion, and less neuroticism. These results provide support for the causal mechanisms running from subjective well-being to later income.

- Adding sibling fixed effects allows for making inferences about lagged effect of well-being at a particular time points instead of having to consider variation between time intervals.

Key Findings

- This paper reverses one of the most famous questions of social science, exploring the influence of income upon well-being.
- Regression equations of happiness on income will produce unreliable results unless the endogeneity of income is incorporated.

Unanswered Questions

- How do we consistently measure "happiness" or "well-being" — this paper seems like it may not reproduce in other contexts because self-reported happiness is highly contextual. Strong interviewer effects, strong seasonal effects, etc.

Goldstein, Joshua R., and Guy Stecklov. 2016.

Citation

Goldstein, Joshua R., and Guy Stecklov. 2016. "From Patrick to John F.: Ethnic Names and Occupational Success in the Last Era of Mass Migration." American Sociological Review 81(1):85–106. doi: 10.1177/0003122415621910.

Key Background

- Despite extensive discussion of assimilation, surprisingly little empirical research examines the consequences of cultural assimilation for economic achievement, in part because (i) culture is difficult to measure and (ii) it is difficult to disentangle measure of cultural assimilation from those of socioeconomic success.
- Names are largely constant over a lifetime, so controlling for socioeconomic gradients in name-giving allows for causal identification of cultural assimilation indicated by the identity aspect of a name.
- Names are important symbols of group affiliation, connoting class and ethnicity. Further, "Black"-sounding names are associated with discrimination in the labor market, as demonstrated by audit studies.
- Names are a proxy measure of a broader set of intentions, cultural markers used by parents to express their own desires and traditions.
- A name is distinctive if it is disproportionately held by members of a particular group.

Research Question

- Are immigrants who assimilated better—as measured by how "American" sounding their name is—able to more effectively climb the economic ladder in the US?

Methods

- Construct a score of the ethnic distinctiveness of names based on relative frequency of each name:

$$ENI(name_j, ethnicity_k) = \frac{p(name_j|ethnicity_k)}{p(name_j|ethnicity_k) + p(name_j|native)} \quad (2.4)$$

where p is the fraction holding given names among j people in ethnicity k.

- Regression of Ethnic Name Index (ENI) on occupational prestige score with vector of controls including age, urban/rural residence, region of residence, and nativity of mother. Second regression with controls for the correlation between name-giving and class background + occupational prestige score of fathers by name of son for father-son pairs.

Argument / contribution

— For most groups, cultural assimilation as measured by first names is strongly positively correlated with occupational achievement. Russian, predominantly Jewish, immigrants are an exception. - Historical datasets with names can be useful for related research topics + interests, including (1) historical effects of distinctively black names, (2) ethnic names on marriage and ethnic intermarriage, (3) the effect of religion on fertility, using biblical names as proxy. - Methods are well-sited for applications to contemporary immigration and naming; further, only requires frequency of names not entire spelling itself (coded numeric string can be used as proxies for distinct names).

Key Findings

- It it possible to translate names into quantitative indicies of cultural assimilation for children of immigrants, and this information can be easily incorporated into models of occupational attainment.
- Controlling for social class gradient allows for a more causal interpretation, where parents who are able to and chose to make their children "more American" confer an advantage in terms of occupation achievement.

Unanswered Questions

Do these research findings also extend to women?

Taubman, Sarah L., James M. Robins, Murray A. Mittleman, and Miguel A. Hernán. 2009.

Citation

Taubman, Sarah L., James M. Robins, Murray A. Mittleman, and Miguel A. Hernán. 2009. "Intervening on Risk Factors for Coronary Heart Disease: An

Application of the Parametric g-Formula." International Journal of Epidemiology 38(6):1599–1611. doi: 10.1093/ije/dyp192.

Key Background

- Epidemiologists often want to estimate the effects of hypothetical interventions to inform policy and clinical decisions. While ideally these questions could be answered with large, randomized experiments, in practice we need to infer answers from observational longitudinal studies. — The g-formula provides a framework to estimate the effects of hypothetical interventions, including joint and dynamic intervention, from complex longitudinal data.

Research Question

- What is the causal effect of different interventions (e.g., avoid smoking, exercise 30 min a day, consuming alcohol) on the 20-year CHD risk?

Methods

- Data from the Nurses' Health Study (N = 121,701) registered nurses aged 30-55 respondent to a mailed questionnaire.
- Employed the parametric g-formula to consistently estimate CHD risk under a hypothetical intervention assuming all joint predictors of the outcome and of the exposures involved in the intervention are measured at all time points.

Argument / contribution

— The G-Formula relies on the same assumptions (no unmeasured confounding, no measurement error, and no model misspecification) as other standard methods, but it correctly adjusts for time-varying confounders affected by prior exposure. - The authors assume no loss to follow-up, and intervene to set C_k = 0 for all time points.

Key Findings

— Despite its limitations, the G-Comp Formula is a powerful and useful tool for epidemiologic analysis, and can be used to compare hypothetical interventions from observational cohort studies under the assumption of no unmeasured confounding. - Hypothetical interventions can be dynamic or static, and can involve multiple exposures.

Unanswered Questions

Chapter 3

Social Networks

We describe our methods in this chapter.