

Design Sensitivity in Observational Studies

Paul R. Rosenbaum, University of Pennsylvania

References

- [1] Rosenbaum, P. R. (2004) Design sensitivity in observational studies. *Biometrika*, 91, 153-164.
- [2] Rosenbaum, P. R. (2005) Heterogeneity and causality: Unit heterogeneity and design sensitivity in observational studies. *American Statistician*, 59, 147-152.
- [3] Small, D. and Rosenbaum, P. R. (2007) War and wages: The strength of instrumental variables and their sensitivity to unobserved biases. *Journal of the American Statistical Association*, to appear.

1 Some terms

Observational studies: (Cochran 1965) Studies of the effects caused by a treatment when ethical or practical issues prevent random assignment to treatment or control, as in a randomized experiment.

Causality: (Neyman 1923 / Rubin 1974) The causal effect of a treatment on a unit compares the response the unit would exhibit under treatment to the response the unit would exhibit under control.

Sensitivity to unobserved bias: Absent random assignment, groups may look similar in terms of observed covariates, but may differ in unobserved covariates. The degree to which causal conclusions change as assumptions about unobserved covariates are changed is the sensitivity of those conclusions to unobserved bias. Less sensitivity is better.

Design sensitivity: The effect of research design on sensitivity to unobserved bias.

2 The basic question (which will be asked repeatedly)

Bias is possible: In any observational study, bias is possible, if not likely. So we adjust for observed covariates, and conduct a sensitivity analysis, indicating whether small or large unobserved biases would alter the conclusions.

If we are fortunate, if there are no unobserved biases in our study, then we would not know this from the observable data.

In this fortunate situation, the best we can hope for is to be able to report that our conclusions are insensitive to small or moderate unobserved biases.

What aspects of the design of an observational study will produce conclusions insensitive to small or moderate unobserved biases when the fortunate situation arises?

Outline

- Informal advice about design to address unobserved biases. What people say.
- Informal application of this informal advice. Does what people say seem correct in context?
- Sensitivity analysis. Are results less sensitive to unobserved bias when the informal advice is followed?
- Aspects of design that affect the power of a sensitivity analysis.
- Design sensitivity. The ability of competing designs to resist unobserved biases in large samples.

3 Informal Advice about Design

Fields: Several fields conduct observational studies of human populations, including economics, epidemiology and medicine, sociology and public program evaluation, psychology and psychiatry.

Advice for design: Diverse advice about design, stated informally, suggests that design matters for unobserved biases. Contrasts with unified advice about experimental design.

Informal advice (talk): May be, often is, good advice. Differences of opinion may be hard to resolve. Qualitative, not quantitative.

Formal advice (theorems): Differences resolvable (in principle). Quantitative appraisal.

4 Sources of some good, informal advice

Hill, A. B. (1965) Environment and disease: association or causation? *Proc Roy Soc Med.*

Shadish, W. R., Cook, T. D. & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.*

Meyer, B. D. (1995). Natural and quasi-experiments in economics. *J. Bus. Econ. Statist.*

Rosenzweig, M. R. & Wolpin, K. I. (2000) Natural “natural experiments” in economics. *J. Econ. Lit.*

Angrist, J. D. & Krueger, A. B. (2001) Instrumental variables and the search for identification. *J. Econ. Perspec.*

Hamermesh, D. S. (2000), “The craft of labormetrics,” *Industrial and Labor Relations Review*

5 Some quoted advice (Dose)

In his President's Address to the Royal Society of Medicine, Sir Austin Bradford Hill (1965) asked:

Our observations reveal an association between two variables, perfectly clear-cut and beyond what we would care to attribute to the play of chance. What aspects of that association should we especially consider before deciding that the most likely interpretation of it is causation?

Hill's causal criteria: Hill's proposed nine considerations including: "biological gradient or dose-response" and "coherence." In most epidemiology texts.

EG: Heavy smokers experience more lung cancer than light smokers.

Does it work? Does a dose-response relationship strengthen causal claims?

6 Disagreements about advice (Dose)

Rothman (1986, p18): “Some causal associations, however, show no apparent trend of effect with dose; an example is the association between DES and adenocarcinoma of the vagina . . . Associations that do show a dose-response trend are not necessarily causal; confounding can result in such a trend between a noncausal risk factor and disease if the confounding factor itself demonstrates a biologic gradient in its relation with disease.”

Weiss (1981, p. 488): Weiss (1981, p. 488): “. . .one or more confounding factors can be related closely enough to both exposure and disease to give rise to [a dose response relationship] in the absence of cause and effect.”

Does it work? Does a dose-response relationship strengthen causal claims? (Could it be a matter of degree? Can the data guide us?)

7 More quoted advice (Coherence)

Cook & Shadish (1994, p. 565) “Successful prediction of a complex pattern of multivariate results often leaves few plausible alternative explanations.”

Trochim (1985, p. 580): “. . . with more pattern specificity it is generally less likely that plausible alternative explanations for the observed effect pattern will be forthcoming.”

Multiple operationalism. Campbell (1988, p. 33):
“. . . great inferential strength is added when each theoretical parameter is exemplified in two or more ways, each mode being as independent as possible of the other, as far as the theoretically irrelevant components are concerned.”

Does it work? Does coherence or pattern specificity or multiple operationalism strengthen causal claims?

8 More quoted advice (Heterogeneity)

In 1864, in his *System of Logic: Principles of Evidence and Methods of Scientific Investigation*, John Stuart Mill proposed “four methods of experimental inquiry,” including the “method of difference:”

If an instance in which the phenomenon . . . occurs and an instance in which it does not . . . have every circumstance save one in common . . . [then] the circumstance [in] which alone the two instances differ is the . . . cause or a necessary part of the cause (III,§8)

Mill wanted: A complete absence of heterogeneity, “have every circumstance save one in common,” that is, on treated and control units that are identical but for the treatment.

9 Fisher disagreed with Mill

Mill wanted: “two instances . . . exactly similar in all circumstances except the one” under study.

Fisher disagreed: In his *Design of Experiments*, discussing the ‘lady tasting tea,’ Fisher (1935, 1949, p. 18) wrote: “It is not sufficient remedy to insist that ‘all the cups must be exactly alike’ in every respect except that to be tested. For this is a totally impossible requirement . . . These are only examples of the differences probably present; it would be impossible to present an exhaustive list of such possible differences . . . because [they] . . . are always strictly innumerable. When any such cause is named, it is usually perceived that, by increased labor and expense, it could be largely eliminated. Too frequently it is assumed that such refinements constitute improvements to the experiment . . .”

Driving out heterogeneity: Was Mill as wrong as Fisher said he was?

10 Instrumental Variables

Instrument: Haphazard event that manipulates a treatment without fully controlling it, and affects the outcome only indirectly by manipulating the treatment. (Angrist's 1990 use of the Vietnam War draft lottery to study the effects of military service on earnings.)

Useful but hard to find: Helps with causal inference, but hard to find, and hard to know when you've found one.

Valid: if IV assumptions are true.

Strong: if the manipulation is strong.

Question: Should we prefer a valid but weak instrument over possibly somewhat biased but stronger instrument?

11 Three examples (one theoretical), used repeatedly

Slightly stylized examples: To focus attention on unobserved biases, control of bias from observed covariates is by forming matched pairs. Mill vs Fisher on heterogeneity viewed through a simulated example.

Genetic toxicology example: Does cigarette smoking cause chromosome damage? (Sierra-Torres, et al. 2004).

Labor economics/IV example: Did World War II military service affect the earnings of U.S Veterans? Adapted from Angrist and Krueger (1994).

Informal view: The premise of the informal advice is that there are things we can see in observable data that are relevant to unobserved biases. Proceeding informally, does that appear to be true?

12 Genetic toxicology example (Sierra-Torres, et al 2004)

Study: 52 smokers, closely matched for age with 52 nonsmokers. (Gender was balanced, not matched.)

Dose of smoking: Pack-years = (packs of cigarettes per day) \times (# years smoked). Strongly correlated with age.

2 Outcomes: For each person, drew a blood sample, established a cell culture, harvested lymphocytes, and for 100 metaphase cells, counted the number of chromatid-type aberrations and the number of chromosome type aberrations.

Aspects: Has both issues of dose-response and coherence.

13 Figure showing types of aberrations

Chromatid is one strand of a pair of chromosomes.

Chromatid aberrations are typically more localized and more common (and perhaps less serious), while chromosome aberrations reflect larger, rarer damage (perhaps more serious).

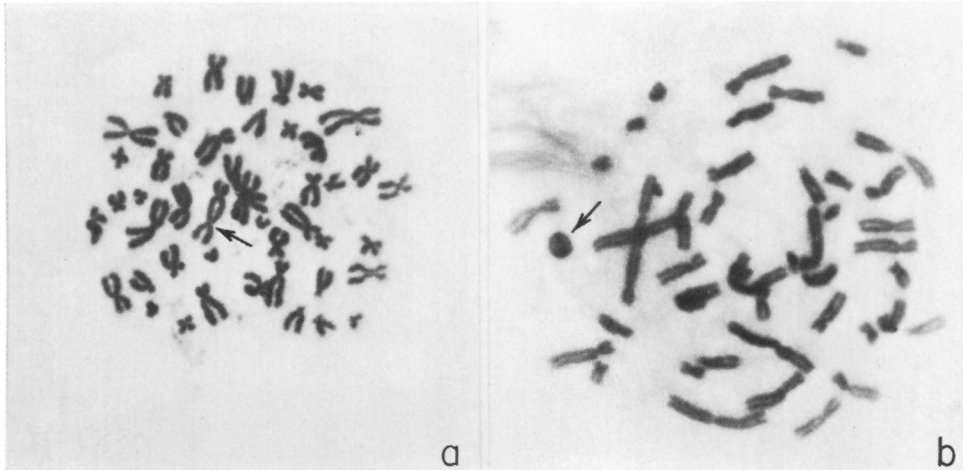


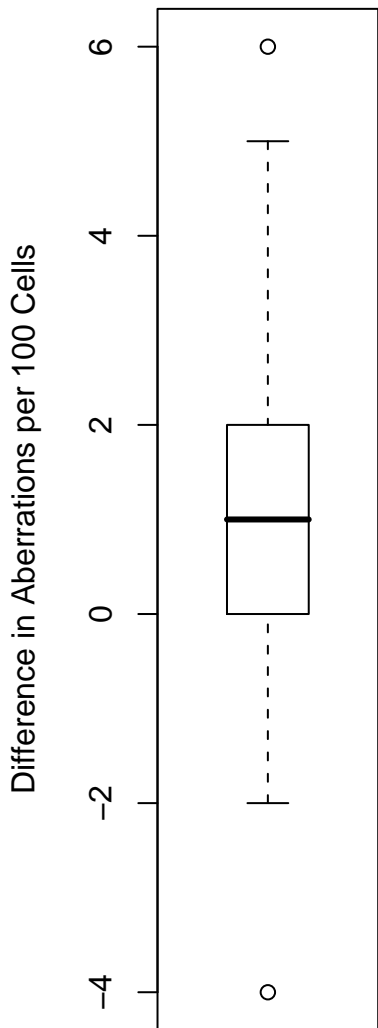
FIG. 1. Aberrant chromosomes of cells from irradiated men. *a*, Cell with dicentric chromosome; *b*, cell with ring chromosome (and also a dicentric).

14 Figure showing coherence & dose-response

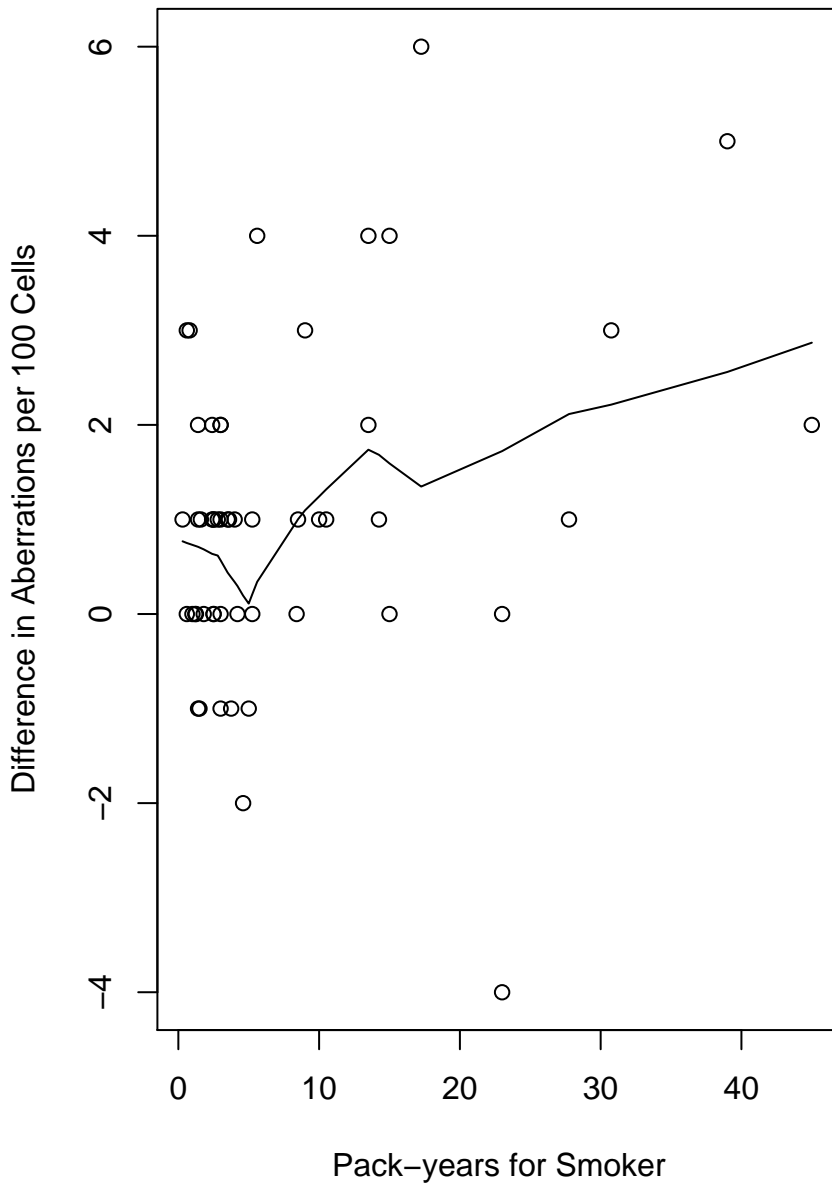
Mention pattern specificity, coherence, dose-response.

Incoherence would have hurt causal claims. Dose what we see help?

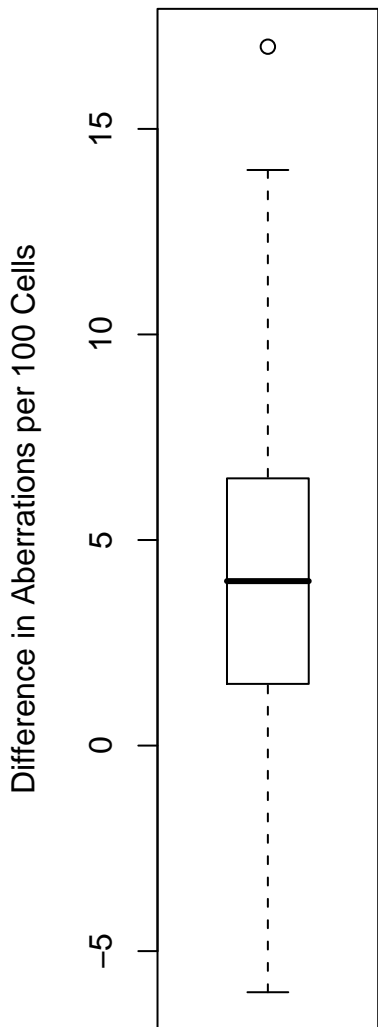
Smoker-Control



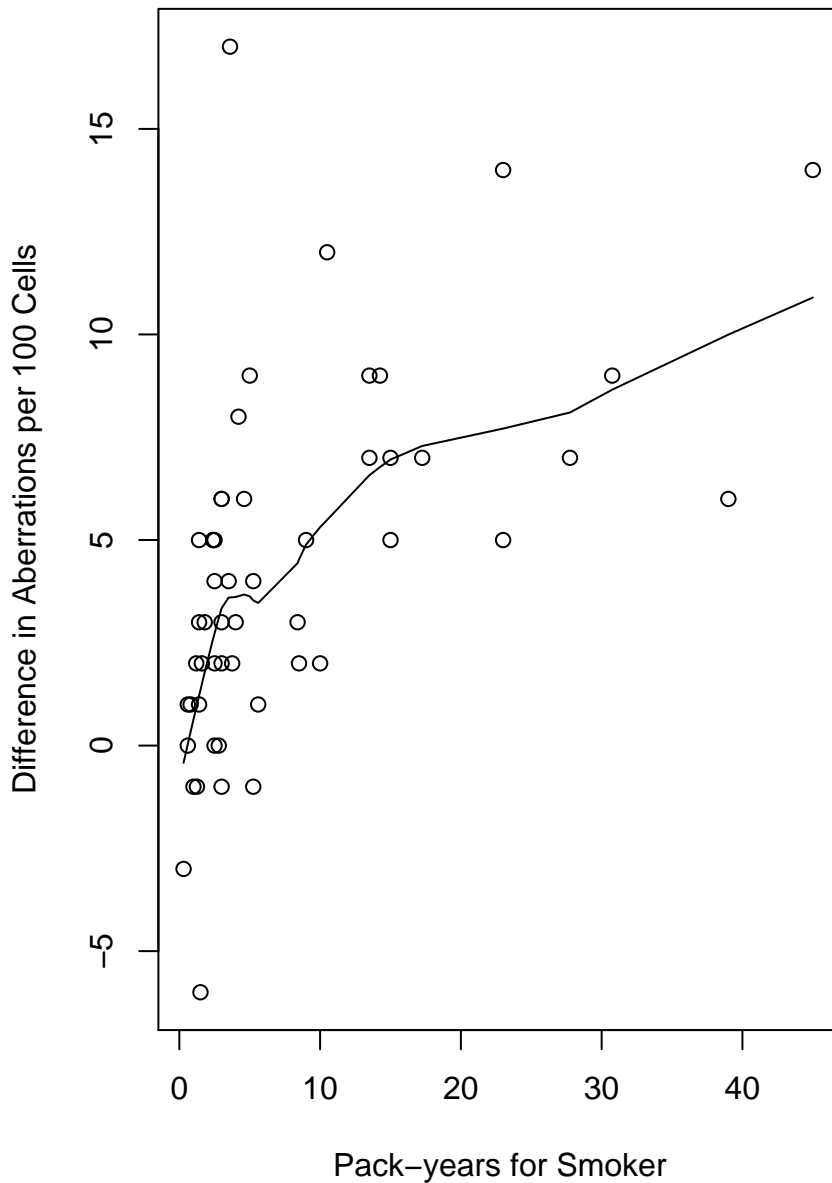
Chromosome Aberrations

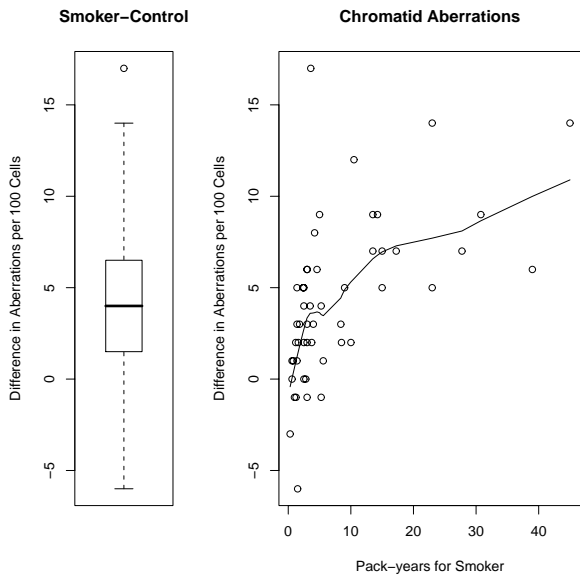


Smoker-Control

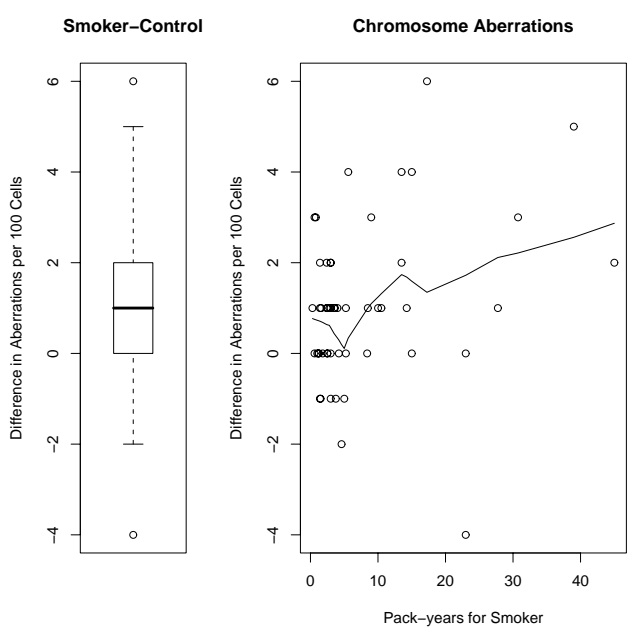


Chromatid Aberrations





Chromatid-type Aberrations in Matched Pairs



Chromosome-type Aberrations in Matched Pairs

15 Labor Economics Example (Adapted from Angrist & Krueger 1994)

Effect of military service on earnings: Service in U.S. military during World War II. Might think service reduces earnings, disrupting education or career. Might think service increases earnings, VA programs, GI Bill, labor market might favor veterans.

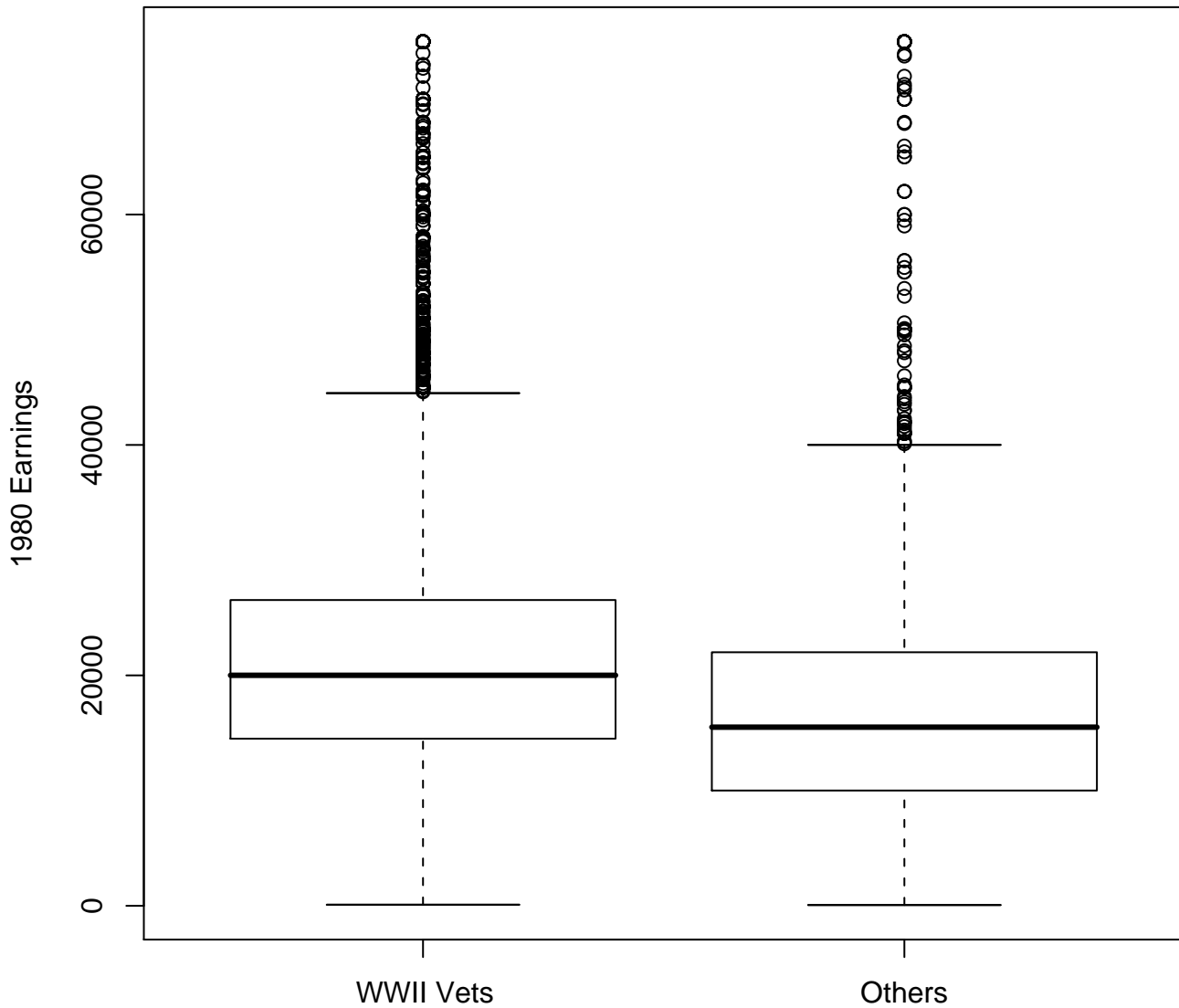
Outcome: Earnings in 1980 obtained from the U.S. Census microdata (5% sample). (Census data have limitations not addressed here or in Angrist & Krueger. Concerns respondents. Demographic data recalled retrospectively. Maximum recorded is US\$75,000.)

A cohort of men: Born in 1926 Q3-Q4, age 18 in 1944, age 54 in 1980. About 76% of this cohort served. Look at 14,000 men. **Vets earned \$4,500 more.**

Picture of Section Bias & IV

16 Selection bias picture

Figure 1: Earnings of Men Born in 1926 Q3 or Q4



17 Two matched cohorts

Cohort early: Born in 1924 Q3-Q4, age 18 in 1942, age 56 in 1980. 78% served.

Cohort late: Born in 1928 Q3-Q4, age 18 in 1946, age 52 in 1980. 24% served.

Matched: 14,000 in each cohort. Matched for (i) quarter of birth, (ii) race (white, black, other), (iii) completed ≥ 8 years of education, (iv) completed ≥ 7 years of education, (v) completed ≥ 6 years of education, (vi) Census region of birth, (vii) Census division of birth, (viii) state of birth. 99% exactly matched for (i) to (vi). 92% exactly matched.

18 IV Picture

Figure 2: Earnings and WWII Veteran Status for Men Born in Q3 or Q4

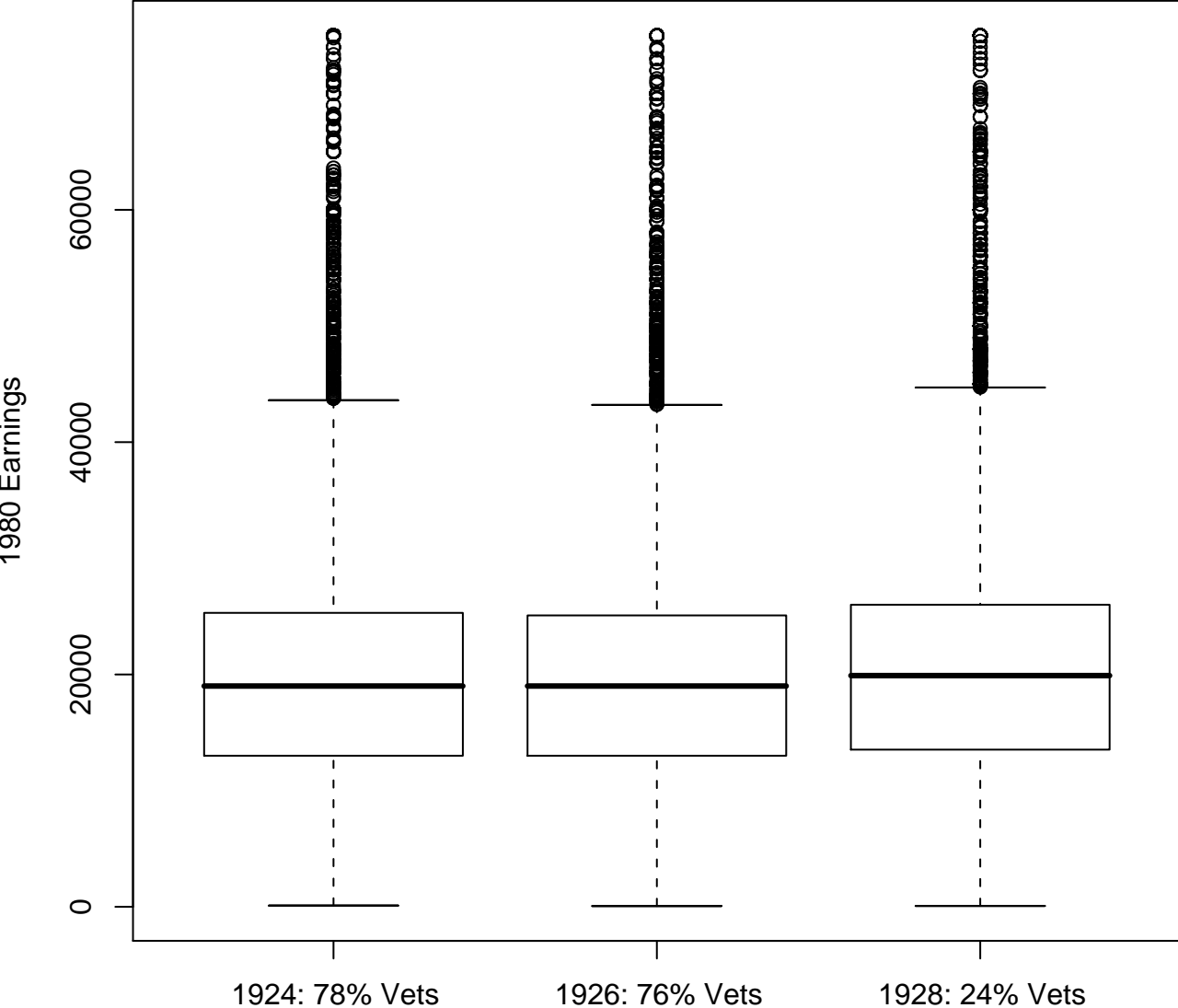


Figure 1: Earnings of Men Born in 1926 Q3 or Q4

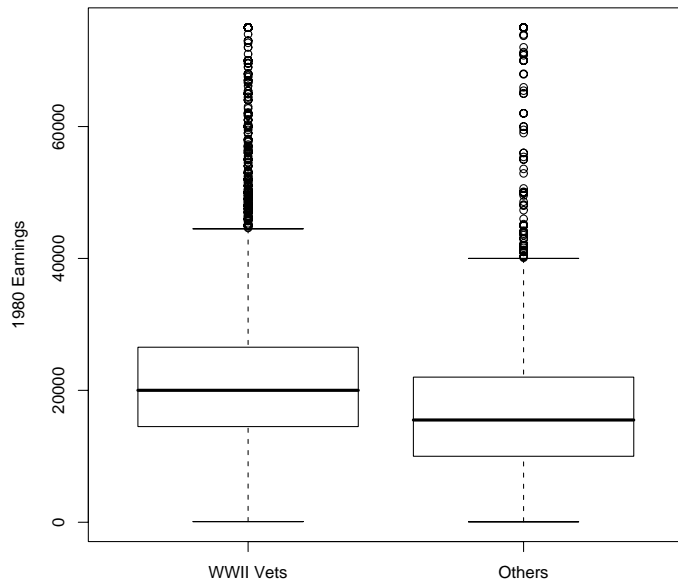
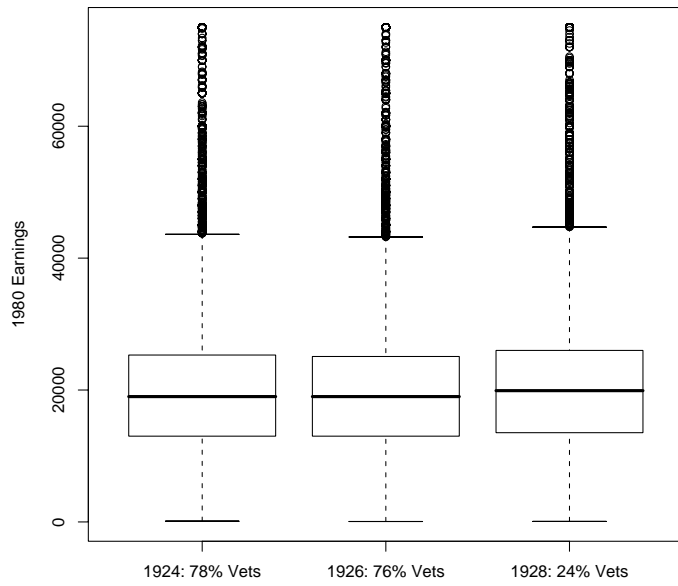
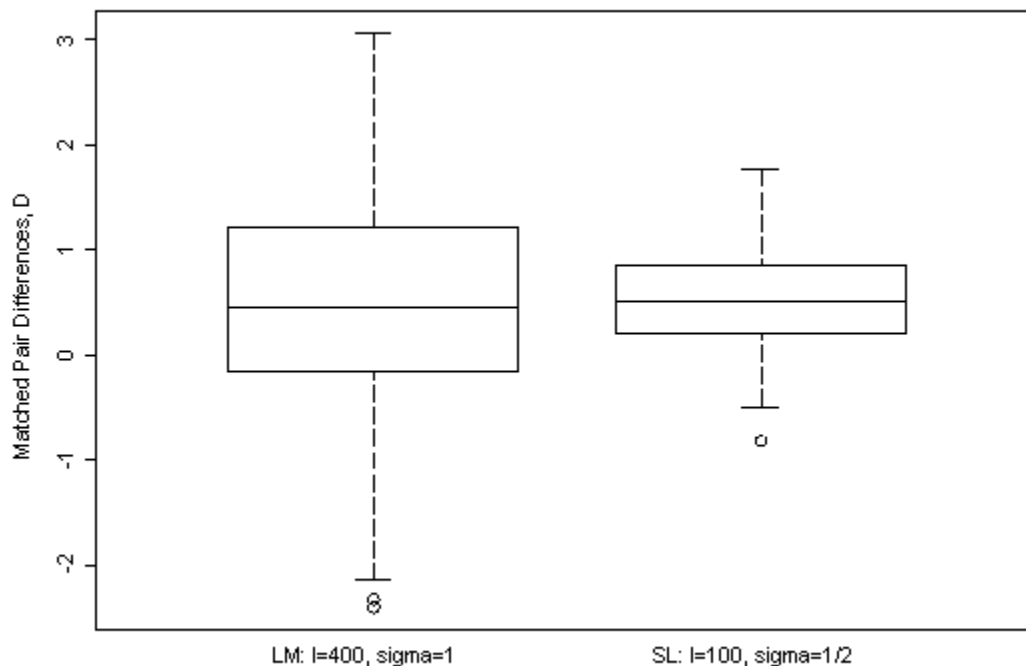


Figure 2: Earnings and WWII Veteran Status for Men Born in Q3 or Q4



19 Mill vs Fisher on Heterogeneity: Theoretical Illustration

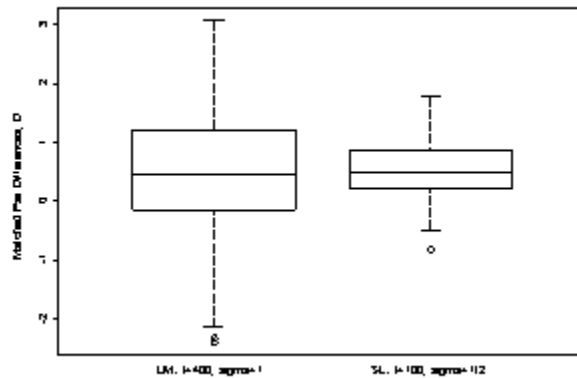
- LM (larger sample size, more heterogeneous): $D_i \stackrel{iid}{\sim} N(\tau, \sigma^2)$ with sample size $4I$ pairs.
- SL (smaller sample size, less heterogeneous): $D_i \stackrel{iid}{\sim} N(\tau, \eta^2)$ with sample size I pairs with $\eta = \sigma/2$.
- Then the sample mean is $\bar{D} \sim N(\tau, \sigma^2/4I)$ in both LM and SL.
- Consider Ashenfelter, O. & Rouse, C. (1998)'s study of the earnings of identical twins with different education. They wanted to remove unobserved biases (which you can't see), but perhaps they simply reduced heterogeneity (which you can see). Is that progress towards causal inference?



Two Simulated Examples. LM has $I = 400$ differences

$$D_i \sim_{iid} N\left(\frac{1}{2}, 1\right). \text{ SL has } I = 100 \text{ differences}$$

$D_i \sim_{iid} N\left\{\frac{1}{2}, \left(\frac{1}{2}\right)^2\right\}$. The randomization inferences are similar, but SL is less sensitive to unobserved bias.



Two Simulated Examples. LM has $I = 400$ differences $D_i \sim iid N\left(\frac{1}{2}, 1\right)$. SL has $I = 100$ differences $D_i \sim iid N\left\{\frac{1}{2}, \left(\frac{1}{2}\right)^2\right\}$. The randomization inferences are similar, but SL is less sensitive to unobserved bias.

	LM	SL
HL estimates $\hat{\tau}$	0.50	0.52
Wilcoxon 95% CI $[\hat{\tau}_L, \hat{\tau}_H]$	[.40, .60]	[.43, .62]

20 Simplest Situation: Matched Pairs

- I pairs, $i = 1, \dots, I$, of two units, $j = 1, 2$, one treated, one control, matched for observed covariates, yielding I treated-minus-control differences in responses, D_i , $i = 1, \dots, I$.
- In the genetic toxicology example, 52 pairs, smoker, nonsmoker, matched for age, and D_i is 2-dimensional, measuring the difference in chromosome damage.
- In the labor economics example, two sets of 14,000 men, paired for race, early education, etc, and D_i is the difference in earnings.
- In a randomized experiment, D_i 's estimate the effect of the treatment and form the basis for randomization inferences about effects (Fisher's 1935).
- In an observational study, D_i 's might tend to be positive even if the treatment had no effect.

21 Notation 1

Covariates: Observed covariate \mathbf{x} . Unobserved covariate u .

Matching: I pairs, $i = 1, \dots, I$, of two subjects, $j = 1, 2$, matched for \mathbf{x} , so $\mathbf{x}_{i1} = \mathbf{x}_{i2}$ for each i , but not for u , so typically $u_{i1} \neq u_{i2}$.

Treatment indicator: $Z_{ij} = 1$ if j received treatment, $Z_{ij} = 0$ if j received control, so $Z_{i1} + Z_{i2} = 1$ for $i = 1, \dots, I$.

Responses: Potential responses, (r_{Tij}, r_{Cij}) , r_{Tij} under treatment, $Z_{ij} = 1$, r_{Cij} under control, $Z_{ij} = 0$, so effect is $r_{Tij} - r_{Cij}$; Neyman (1935) & Rubin (1974).

Doses: May have doses levels of treatment, (v_{Tij}, v_{Cij}) . EG, $v_{Tij} = \text{pack} - \text{years}$, $v_{Cij} = 0$. Without doses, take $v_{Tij} = 1$, $v_{Cij} = 0$.

22 Paired Randomized Experiment

Conditioning: Write

$$\mathcal{F} = \left\{ \left(r_{Tij}, r_{Cij}, v_{Tij}, v_{Cij}, \mathbf{x}_{ij}, u_{ij} \right), \right. \\ \left. i = 1, \dots, I, j = 1, 2 \right\}$$

$$\mathcal{Z} = \{ Z_{i1} + Z_{i2} = 1, i = 1, \dots, I \};$$

then \mathcal{F} and \mathcal{Z} are fixed by conditioning in Fisher's theory of randomization inference.

Randomization: $\Pr(Z_{i1} = 1 \mid \mathcal{Z}, \mathcal{F}) = \frac{1}{2}, i = 1, \dots, I,$
with independent assignments in distinct pairs.

Observed responses, differences: R_{ij} observed is $R_{ij} = Z_{ij} r_{Tij} + (1 - Z_{ij}) r_{Cij}$, and the treated-minus-control difference in responses in pair i is $D_i = (2Z_{i1} - 1) (R_{i1} - R_{i2})$.

23 Common Effect Models in a Paired Randomized Experiment

Constant effect: If the treatment effect is constant, $\tau = r_{Tij} - r_{Cij}$, then $R_{ij} = r_{Cij} + Z_{ij}\tau$, and $D_i = (2Z_{i1} - 1) (R_{i1} - R_{i2}) = \tau + \epsilon_i$ where $\epsilon_i = (2Z_{i1} - 1) (r_{Ci1} - r_{Ci2})$.

Effect proportional to dose: If

$$r_{Tij} - r_{Cij} = \beta (v_{Tij} - v_{Cij}),$$

which is constant effect if $v_{Tij} = 1, v_{Cij} = 0$.

24 Wilcoxon's Signed Rank Statistic

Wilcoxon's Signed Rank Statistic: To test $H_0 : \tau = \tau_0$ rank $|D_i - \tau_0|$ from 1 to I ; then W_{τ_0} , is the sum of the ranks for which $D_i - \tau_0 > 0$, where ties are assumed absent.

As a randomization test: If $H_0 : \tau = \tau_0$ is true, randomization ensures $D_i - \tau_0 = \epsilon_i$ is $r_{Ci1} - r_{Ci2}$ or $r_{Ci2} - r_{Ci1}$, each with probability $\frac{1}{2}$, independently in different pairs. Given \mathcal{Z}, \mathcal{F} , if $H_0 : \tau = \tau_0$ is true, then W_{τ_0} is the sum of I independent random variables taking values i or 0 each with probability $\frac{1}{2}$, $i = 1, \dots, I$.

Confidence interval: A confidence interval for τ is obtained by inverting the test.

HL estimate: Hodges-Lehmann (1963) or HL estimate of τ is (essentially) the value $\hat{\tau}$ such that $W_{\hat{\tau}}$ is as close as possible to its null expectation, $I(I + 1) / 4$.

25 Models and Power

So far, just randomization inference: The null distribution of W_{τ_0} is the same for all (untied) \mathcal{F} . All that was used for test, CI and estimate.

Power: The nonnull distribution of W_{τ_0} depends on \mathcal{F} or a model that generates \mathcal{F} .

Common model for power: $(r_{Ci1} - r_{Ci2}) / \sigma \sim_{iid} F(\cdot)$
where $\sigma > 0$ and $F(\cdot)$ is a continuous distribution symmetric about zero, so that randomization ensures $\epsilon_i / \sigma \sim_{iid} F(\cdot)$.

Reference: Lehmann (1998, §3-§4).

26 Departures from Random Assignment

1. In the population prior to matching, treatment assignments were independent, with unknown probabilities $\pi_{ij} = \Pr(Z_{ij} = 1 \mid \mathcal{F})$
2. Two subjects with the same *observed* \mathbf{x}_{ij} may differ in *unobserved* u_{ij} and hence in their odds of receiving treatment by a factor of $\Gamma \geq 1$,

$$\frac{1}{\Gamma} \leq \frac{\pi_{ij}(1 - \pi_{ik})}{\pi_{ik}(1 - \pi_{ij})} \leq \Gamma, \quad \forall i, j, k \quad (1)$$

3. Distribution of treatments within treated/control matched pairs $\Pr(Z_{i1} = 1 \mid \mathcal{Z}, \mathcal{F})$ is then obtained by conditioning on $Z_{i1} + Z_{i2} = 1$.

27 Departures, continued

$$1/\Gamma \leq \left\{ \pi_{ij} (1 - \pi_{ik}) \right\} / \left\{ \pi_{ik} (1 - \pi_{ij}) \right\} \leq \Gamma, \quad \mathbf{x}_{ij} = \mathbf{x}_{ik}$$

No unobserved bias: If $\Gamma = 1$, then $\mathbf{x}_{ij} = \mathbf{x}_{ik}$ ensures $\pi_{ij} = \pi_{ik}$, $i = 1, \dots, I$, whereupon

$$\Pr(Z_{i1} = 1 \mid \mathcal{Z}, \mathcal{F}) = \pi_{i1} / (\pi_{i1} + \pi_{i2}) = \frac{1}{2}.$$

Uncertainty from unobserved bias: If $\Gamma > 1$ in (1), then matching on \mathbf{x} may fail to equalize the π_{ij} in pair i , and $\Pr(Z_{i1} = 1 \mid \mathcal{Z}, \mathcal{F})$ is unknown.

Question answered by a sensitivity analysis: Bounds on significance levels, point estimates, confidence intervals for several values of Γ . How large must Γ be before qualitatively different causal interpretations are possible?

28 Sensitivity Analysis Procedure

Two known distributions: For fixed $\Gamma \geq 1$, let $\overline{\overline{W}}$ be the sum of I independent random variables taking value i with probability $\theta = \Gamma / (1 + \Gamma)$ and value 0 with probability $1 - \theta$, $i = 1, \dots, I$; and let \overline{W} for the sum of I independent random variables taking value i with probability $1 - \theta$ and value 0 with probability θ .

Bounds: If

$$\frac{1}{\Gamma} \leq \frac{\pi_{ij} (1 - \pi_{ik})}{\pi_{ik} (1 - \pi_{ij})} \leq \Gamma, \quad \forall i, j, k$$

and $H_0 : \tau = \tau_0$ are true, then the following bounds are sharp for each $\Gamma \geq 1$:

$$\Pr(\overline{W} \geq w) \leq \Pr(W_{\tau_0} \geq w \mid \mathcal{Z}, \mathcal{F}) \leq \Pr(\overline{\overline{W}} \geq w)$$

Cases: If $\Gamma = 1$, then equality; otherwise the bounds become wider as Γ increases.

29 Procedure, continued

For each $\Gamma \geq 1$

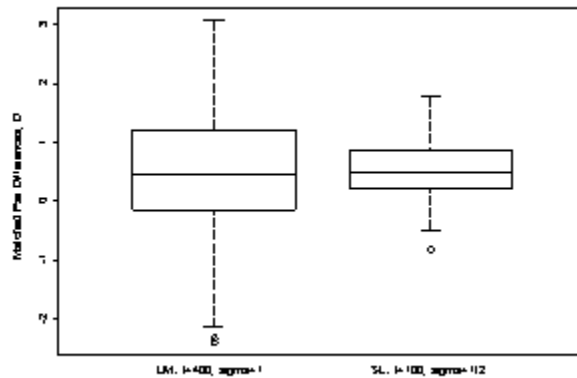
$$\Pr(\overline{W} \geq w) \leq \Pr(W_{\tau_0} \geq w \mid \mathcal{Z}, \mathcal{F}) \leq \Pr(\overline{\overline{W}} \geq w)$$

Tests: For each $\Gamma \geq 1$, test $H_0 : \tau = \tau_0$ versus $H_A : \tau > \tau_0$; then the upper bound on the one-sided significance level is at most 0.05 for all π_{ij} if $W_{\tau_0} \geq \tilde{w}$ where $0.05 = \Pr(\overline{\overline{W}} \geq \tilde{w})$.

Estimates: Recall that $\theta = \Gamma / (1 + \Gamma)$. Bounds on the expectation of W_τ

$$\frac{(1 - \theta) I (I + 1)}{2} \leq E(W_\tau \mid \mathcal{Z}, \mathcal{F}) \leq \frac{\theta I (I + 1)}{2}$$

yield an interval of HL point estimates, $[\hat{\tau}_{\min}, \hat{\tau}_{\max}]$. With no unobserved bias, $\Gamma = 1$, $\mu_\pi = I(I + 1)/4$, and $\hat{\tau}_{\min} = \hat{\tau}_{\max}$ is the usual HL estimate.



Two Simulated Examples. LM has $I = 400$ differences $D_i \sim iid N\left(\frac{1}{2}, 1\right)$. SL has $I = 100$ differences $D_i \sim iid N\left\{\frac{1}{2}, \left(\frac{1}{2}\right)^2\right\}$. The randomization inferences are similar, but SL is less sensitive to unobserved bias.

	LM	SL
HL estimates $\hat{\tau}$	0.50	0.52
Wilcoxon 95% CI $[\hat{\tau}_L, \hat{\tau}_H]$	[.40, .60]	[.43, .62]

Sensitivity Analysis for Testing

$$H_0 : \tau = 0 \text{ vs } H_A : \tau > 0$$

Values are upper bounds on one-sided significance levels.

	LM	SL
$\Gamma = 1$	10^{-14}	10^{-14}
$\Gamma = 2$	0.00037	0.00000032
$\Gamma = 3$	0.37	0.000088
$\Gamma = 5$	1.00	0.0078

HL Estimates

		LM	SL
$\Gamma = 1$	$\hat{\tau} = \hat{\tau}_{\min} = \hat{\tau}_{\max}$	0.50	0.52
$\Gamma = 2$	$[\hat{\tau}_{\min}, \hat{\tau}_{\max}]$	[.19, .81]	[.37, .67]

30 Power of a sensitivity analysis

- The sensitivity analysis reported a sharp upper bound on the one-sided *p* – *value* testing $H_0 : \tau = 0$ vs $H_A : \tau > 0$.
- The power of a sensitivity analysis is the probability, under some alternative, that this upper bound is less than 0.05.
- The alternative considered here is:
 1. The treatment worked, with constant effect $\tau = \frac{1}{2}$. (But of course, we don't know this.)
 2. We were fortunate, and there was no unobserved bias, $\Gamma = 1$. (But of course, we don't know this.)
 3. Errors are Normal, Logistic or Cauchy. (But of course, we don't know this.)

Power of the Sensitivity Analysis: Normal Errors

(treatment effect $\tau = \frac{1}{2}$, no unobserved bias)

	LM	SL
I pairs	120	30
σ	1	$\frac{1}{2}$
$\Gamma = 1$	1.00	1.00
$\Gamma = 1.5$	0.96	1.00
$\Gamma = 2$	0.60	0.96

Power of the Sensitivity Analysis: Logistic Errors

(treatment effect $\tau = \frac{1}{2}$, no unobserved bias)

	LM	SL
I pairs	120	30
σ	1	$\frac{1}{2}$
$\Gamma = 1$	0.93	0.93
$\Gamma = 1.5$	0.31	0.61
$\Gamma = 2$	0.04	0.32

Power of the Sensitivity Analysis: Cauchy Errors

(treatment effect $\tau = \frac{1}{2}$, no unobserved bias)

	LM	SL
I pairs	200	50
σ	1	$\frac{1}{2}$
$\Gamma = 1$	0.98	0.95
$\Gamma = 1.5$	0.32	0.60
$\Gamma = 2$	0.02	0.28

31 Limiting case, $I \rightarrow \infty$

- As the number of pairs $I \rightarrow \infty$, the only uncertainty that remains is due to unobserved bias.
- In particular, for each $\Gamma \geq 1$, as $I \rightarrow \infty$, the (random) interval of HL estimates, $[\hat{\tau}_{\min}, \hat{\tau}_{\max}]$ converges in probability to a fixed interval $[\tau_{\min}, \tau_{\max}]$.
- If $\Gamma = 1$, then $\tau_{\min} = \tau_{\max} = \tau$.
- If $\Gamma > 1$, then $\tau_{\min} < \tau_{\max}$, with $\tau \in [\tau_{\min}, \tau_{\max}]$.

32 Limiting case, $I \rightarrow \infty$

Notation: $\Phi(\cdot)$ and $\Upsilon(\cdot)$ are standard Normal and Cauchy cumulative distributions. Also, $\theta = \Gamma / (1 + \Gamma)$.

Situation: Unknown to us, there actually is no unobserved bias.

Question: For fixed Γ in the sensitivity analysis, how does unit heterogeneity σ affect the limiting interval $[\tau_{\min}, \tau_{\max}]$?

Proposition: If $(D_i - \tau) / \sigma \stackrel{iid}{\sim} \Phi(\cdot)$ then

$$[\tau_{\min}, \tau_{\max}] = \tau \pm \frac{\sigma \Phi^{-1}(\theta)}{\sqrt{2}}$$

If $(D_i - \tau) / \sigma \stackrel{iid}{\sim} \Upsilon(\cdot)$ then

$$[\tau_{\min}, \tau_{\max}] = \tau \pm \sigma \Upsilon^{-1}(\theta).$$

33 Theoretical Point

LM:

$$\frac{D_i - \tau}{\sigma} \stackrel{iid}{\sim} F(\cdot), \quad i = 1, \dots, 4I$$

SL:

$$\frac{D_i - \tau}{\sigma/2} \stackrel{iid}{\sim} F(\cdot), \quad i = 1, \dots, I$$

In a randomized experiment: Not a big difference.

In an observational study: SL much better — less sensitive to unobserved biases.

34 Genetic Toxicology Example

- 52 pairs, matched for age, two measures of chromosome damage. Two Wilcoxon statistics, $W^{(1)}$, $W^{(2)}$, one based on chromatid aberrations, the other on chromosome aberrations, testing the null hypothesis of no effect of smoking.
- Interval of possible significance levels. Interval is a single significance level for $\Gamma = 1$. In this case, all intervals include some very small significance levels, so only the upper endpoint of the interval is given.

Γ	Chromatid	Chromosome
1	2.5×10^{-8}	2.9×10^{-5}
2	0.00018	0.016
2.5	0.0011	0.052
5	0.043	0.478
5.3	0.053	0.529

35 Do Coherence and Doses Matter for Sensitivity to Bias?

- When higher responses are anticipated for both of two outcomes, signed ranks statistics $W^{(1)}$ and $W^{(2)}$, the coherent signed rank statistic is simply $W^{(1)} + W^{(2)}$. A version with doses weights the absolute ranks by the ranks of the doses, akin to Spearman's rank correlation.
- The sensitivity analysis is similar.
- Does the does-response relationship and the coherent pattern of responses reduce sensitivity to unobserved bias?

36 Genetic Toxicology Example, Continued

Γ	Chromatid	Chromosome	Coherent	w/Doses
1	$\sim 10^{-8}$	$\sim 10^{-5}$	$\sim 10^{-8}$	$\sim 10^{-7}$
2	0.00018	0.016	0.00018	0.00052
2.5	0.0011	0.052	0.0010	0.0020
5	0.043	0.478	0.039	0.033
5.3	0.053	0.529	0.048	0.039
5.8	0.072	0.607	0.064	0.050

In this one example, coherence among responses and a dose response relationship measurably reduced sensitivity to unobserved biases.

37 Return to the Original Question

Question: Is what happened in the example to be expected in general?

The fortunate situation: Suppose that our study is actually not affected by unobserved biases. In this situation, what aspects of design would yield a high degree of insensitivity to unobserved biases?

In the fortunate situation, the data (\mathcal{F} and \mathbf{Z}) are generated by some process free of unobserved biases. If a sensitivity analysis were applied to data from this process, what would happen?

In each sample, the sensitivity analysis would yield a value of Γ such that the conclusions begin to change, to become sensitive at that Γ . Varies with data and I .

38 Design Sensitivity: A Parameter Describing a Study Design

In large samples, the situation is simpler. The design sensitivity is a parameter somewhat akin to Pitman efficiency. It compares competing methods/designs for the same problem when the sample size is large.

In the fortunate situation in large samples, there is a number, $\tilde{\Gamma}$, called the design sensitivity, such that the power of a sensitivity analysis tends, as $I \rightarrow \infty$, either to 1 for $\Gamma < \tilde{\Gamma}$ and to 0 for $\Gamma > \tilde{\Gamma}$.

In words, no matter how large the study becomes, results will be sensitive to biases larger than $\tilde{\Gamma}$, and for sufficiently large samples, will be insensitive to biases smaller than $\tilde{\Gamma}$.

39 Intuition behind the mechanics

We have a statistic, say \widetilde{W} , suitably normalized, and in the fortunate situation with no unobserved biases, $\sqrt{I} (\widetilde{W} - \mu) / \omega \rightarrow_D N(0, 1)$, as $I \rightarrow \infty$. For Wilcoxon's signed rank statistic, $\widetilde{W} = 2W / \{I(I + 1)\}$, say.

Although we are in the fortunate situation, we don't know this from the data and perform a sensitivity analysis.

In the absence of a treatment effect, for a given Γ in the sensitivity analysis, the largest possible distribution of \widetilde{W} (e.g., the distribution of $\sqrt{I} 2\overline{\widetilde{W}} / \{I(I + 1)\}$) tends to $N(\overline{\mu}_\Gamma, \overline{\omega}_\Gamma^2)$.

Design sensitivity is found as the solution $\widetilde{\Gamma}$ in Γ to the equation $\mu = \overline{\mu}_\Gamma$. Depends only on limiting expectations.

40 Example of mechanics: Wilcoxon's Signed Rank Statistic

In the fortunate situation, without unobserved biases, additive effect τ , and $H_0 : \tau = \tau_0$ true, $D_i = \tau + \epsilon_i$, with ϵ_i iid, continuous, symmetric about zero,

$$E(W_{\tau_0}) = \frac{I(I-1) \Pr(\epsilon_1 + \epsilon_2 > 0)}{2} + I \Pr(\epsilon_1 > 0).$$

The upper bounding distribution of $\overline{\overline{W}}$ in the sensitivity analysis has

$$E(\overline{\overline{W}}) = \frac{\Gamma}{1 + \Gamma} \frac{I(I+1)}{2}$$

Solve for design sensitivity: Divide by I^2 , let $I \rightarrow \infty$, and solve to get

$$\tilde{\Gamma} = \frac{\Pr(\epsilon_1 + \epsilon_2 > 0)}{1 - \Pr(\epsilon_1 + \epsilon_2 > 0)}$$

41 Return to question about dose-response and coherence

I Matched sets: In each, one treated person with dose matched v_i matched to k untreated controls with dose zero. Will consider $k = 5$.

p Coherent outcomes: Each having a linear regression on dose, with the same slope $\beta = \frac{1}{2}$, and symmetrically correlated multivariate Normal errors, with variances 1 and intercorrelations ρ . Will consider $p = 1$ and $p = 3$ outcomes and correlations $\rho = 0$ and $\rho = \frac{1}{2}$.

Three dose levels: each with a third of treated subjects. Will consider three possible patterns: $(\frac{1}{2}, 1, \frac{3}{2})$, $(1, 1, 1)$, $(\frac{3}{2}, \frac{3}{2}, \frac{3}{2})$.

Statistic: Stratified coherent Wilcoxon rank sum weighted by doses. (Details and other cases in 2004 paper.)

42 Case of $k = 5$ Controls Per Set

- Design sensitivity, $\tilde{\Gamma}$: p outcomes with correlation $\rho = 0$ or $\frac{1}{2}$ for 3 dose patterns.

$k = 5$		$\rho = 0$	$\rho = \frac{1}{2}$
Doses	p		
$(\frac{1}{2}, 1, \frac{3}{2})$	1	3.0	3.0
	3	6.4	3.8
$(1, 1, 1)$	1	2.6	2.6
	3	5.1	3.2
$(\frac{3}{2}, \frac{3}{2}, \frac{3}{2})$	1	4.1	4.1
	3	11.7	5.6

- $\tilde{\Gamma}$ varies. Coherence among $p = 3$ outcomes has a big impact when $\rho = 0$, but smaller for $\rho = \frac{1}{2}$. (No gain for $\rho = 1$.)
- Dose response, $(\frac{1}{2}, 1, \frac{3}{2})$ vs $(1, 1, 1)$ helps a little, but higher uniform doses $(\frac{3}{2}, \frac{3}{2}, \frac{3}{2})$ are much better.

43 Case of $k = 2$ Controls Per Set

- Value of the design sensitivity, $\tilde{\Gamma}$.
- $k = 2$ controls, p coherent outcomes ($p = 0$ or 3) with symmetric intercorrelation $\rho = 0$ or $\frac{1}{2}$ for 3 different dose patterns.

$k = 2$		$\rho = 0$	$\rho = \frac{1}{2}$
Doses	p		
$(\frac{1}{2}, 1, \frac{3}{2})$	1	2.4	2.4
	3	4.2	2.9
$(1, 1, 1)$	1	2.2	2.2
	3	3.6	2.5
$(\frac{3}{2}, \frac{3}{2}, \frac{3}{2})$	1	3.0	3.0
	3	6.4	3.8

- Same pattern as $k = 5$, but $\tilde{\Gamma}$ is smaller.

44 Power of the sensitivity analysis

- Example, $I = 200$ matched sets, one treated and $k = 3$ controls each, p Gaussian outcomes, intercorrelation ρ , slope $\frac{1}{2}$ with dose for all outcomes, pretty much as before.

Power of the sensitivity analysis when performed with
 $\Gamma = 2$

$k = 3$		$\rho = 0$	$\rho = \frac{1}{2}$
Doses	p		
$(\frac{1}{2}, 1, \frac{3}{2})$	1	0.54	0.54
	3	1.00	0.92
$(1, 1, 1)$	1	0.28	0.28
	3	1.00	0.73
$(\frac{3}{2}, \frac{3}{2}, \frac{3}{2})$	1	0.98	0.98
	3	1.00	1.00

- Ranges from 0.28 to 1.00. Dose response, coherence and larger doses all substantially boast power.

45 Recall Instrumental Variables Example

A cohort of men: Born in 1926 Q3-Q4, age 18 in 1944, age 54 in 1980. About 76% of this cohort served in WWII. Look at 14,000 men.

Cohort early: Born in 1924 Q3-Q4, age 18 in 1942, age 56 in 1980. 78% served.

Cohort late: Born in 1928 Q3-Q4, age 18 in 1946, age 52 in 1980. 24% served.

Matched pairs: Two sets of 14,000 matched pairs. 1926-1924 pairs. 1928-1926 pairs.

Idea: Men who served in WWII are likely to be different from those who did not. However, men born in 1926 are not likely to be very different than those born in 1928. Use birth year as an instrument for service. Check on age trend with 1924

46 Table of Earnings

The 1926 Q3-Q4 cohort. Earnings in US\$ in 1980 of 14,000 Men Born in the Second Half of 1926 by World War II Veteran Status.

	Count	Lower	Median	Upper	Trimean
WWII Vets	10,571	14,510	20,010	26,540	20,268
Others	3,429	10,010	15,510	22,010	15,760
Difference		4,500	4,500	4,530	4,508

- A *naive* estimate of the effects of WWII service on earnings is a *benefit* of \$4,500.

47 Imagining Being Born 2 Years Later

Covariates: Observed covariate \mathbf{x} , race, early education, region of birth. Unobserved covariate u . I pairs, $i = 1, \dots, I = 14,000$, of two men, $j = 1, 2$, matched for \mathbf{x} , not u .

Treatment indicator: $Z_{ij} = 1$ if j born in 1926, $Z_{ij} = 0$ if j born in 1928, so $Z_{i1} + Z_{i2} = 1$ for $i = 1, \dots, I$.

Responses: Potential earnings in 1980, (r_{Tij}, r_{Cij}) , r_{Tij} if born in 1926, $Z_{ij} = 1$, r_{Cij} if born in 1928, $Z_{ij} = 0$. Observe $R_{ij} = Z_{ij}r_{Tij} + (1 - Z_{ij})r_{Cij}$.

Doses: Doses of military service in WWII, (v_{Tij}, v_{Cij}) , with $v_{Tij} = 1$ if would have served if born in 1926, $v_{Tij} = 0$ otherwise, $v_{Cij} = 1$ if would have served if born in 1928, $v_{Cij} = 0$ otherwise. Observe $V_{ij} = Z_{ij}v_{Tij} + (1 - Z_{ij})v_{Cij}$.

48 Permutation Inference for IV

Effect proportional to dose model: Embodies exclusion restriction and says birth year affects earning to the extent that it alters military service.

$$r_{Tij} - r_{Cij} = \beta (v_{Tij} - v_{Cij})$$

If true, then

$$R_{ij} - \beta V_{ij} = r_{Tij} - \beta v_{Tij} = r_{Cij} - \beta v_{Cij} = a_{ij}, \text{ say,}$$

so if $H_0 : \beta = \beta_0$ were true, $R_{ij} - \beta_0 V_{ij} = R_{ij} - \beta V_{ij}$ would be the same whether you were born in 1926 or 1928.

Apply Wilcoxon's signed rank test to: $R_{ij} - \beta_0 V_{ij}$ to obtain confidence intervals and point estimates for β . Based on assuming birth year is random, rather than assume WWII military service is random. Do sensitivity analysis as before.

Upper Bound on One-Sided p-value, 14,000 1928 vs
1926 Pairs, $H_0 : \beta = \beta_0$ vs $H_0 : \beta < \beta_0$.

Γ	$\beta_0 = 0$	$\beta_0 = 1,000$	$\beta_0 = 4,500$	$\beta_0 = 10,000$
1	0.001	0.001	0.001	0.001
1.2	1.000	0.860	0.001	0.001
1.5	1.000	1.000	0.027	0.001
1.6	1.000	1.000	0.904	0.001
2.2	1.000	1.000	1.000	0.016
2.3	1.000	1.000	1.000	0.476

If the instrument were valid: ($\Gamma = 1$) then confident military service reduced earnings. $[-\$1,445, -\$500]$ is 95% CI for β .

Even if the instrument were invalid: if we had left out of the matching a variable u strongly associated with earnings in 1980 and $\Gamma = 1.5$ times more common among men born in 1928 than in 1926, we would still reject the naive \$4,500 benefit as too large with a $p < 0.027$.

49 Weak Instrument

1924 vs 1926, parallel but very weak: In parallel, the small difference in WWII service between the 1924 and 1926 matched cohorts (78% vs 76%) creates a very weak instrument, with a parallel analysis.

Even if perfectly valid ($\Gamma = 1$), the weak instrument provides little information, with a 95% CI for effect β of $[-\$10,130, \$10,750]$ in a population whose median annual earnings were roughly \$20,000.

50 Technical Details

$$\begin{aligned} D_i^{\beta_0} &= Z_{i1} \{(R_{i1} - \beta_0 V_{i1}) - (R_{i2} - \beta_0 V_{i2})\} \\ &\quad + (1 - Z_{i1}) \{(R_{i2} - \beta_0 V_{i2}) - (R_{i1} - \beta_0 V_{i1})\} \\ &= (\beta - \beta_0) \{Z_{i1} (v_{Ti1} - v_{Ci2}) + (1 - Z_{i1}) (v_{Ti2} - v_{Ci1})\} \\ &\quad + (2Z_{i1} - 1) (a_{i1} - a_{i2}) \\ &= (\beta - \beta_0) S_i + \epsilon_i \end{aligned}$$

where

$$S_i = Z_{i1} (v_{Ti1} - v_{Ci2}) + (1 - Z_{i1}) (v_{Ti2} - v_{Ci1})$$

and

$$\epsilon_i = (2Z_{i1} - 1) (a_{i1} - a_{i2})$$

is $\pm (a_{i1} - a_{i2})$ if $H_0 : \beta = \beta_0$ is true.

51 Design Sensitivity Under a Very Simple Model

Compliance: Population has π_A “always takers” with $v_{Ti1} = v_{Ci2} = 1$ who would serve whether born in 1926 or 1928, π_N “never takers” with $v_{Ti1} = v_{Ci2} = 0$ who would not serve whether born in 1926 nor in 1928, and π_C “compliers” with $v_{Ti1} = 1$, $v_{Ci2} = 0$ who would serve only if born in 1926, not if born in 1928. For simplicity of presentation, will consider $\pi_A = \pi_N$, but calculations are equally easy in all cases.

Matched Pair Errors: $\epsilon_i \sim iid$ Normal, Cauchy or logistic, centered at 0, with scale σ , independent of compliance. Effect size measured by $\lambda = (\beta_0 - \beta) / \sigma$.

Examples: $(\pi_A, \pi_C, \pi_N) = (0, 1, 0)$ is perfect compliance. The 1926-1928 pairs look more like 50% compliance, $(\pi_A, \pi_C, \pi_N) = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$. A weak instrument might have 10% compliance, $(\pi_A, \pi_C, \pi_N) = (\frac{9}{20}, \frac{2}{20}, \frac{9}{20})$.

52 Recall Basic Question

- Suppose the instrument were valid. We would not know this from the data.
- We would not know this from the data. So we would appraise sensitivity of conclusions to unobserved biases in the instrument.
- The best we could hope to say is that the results are insensitive to small or moderate biases.
- The power of the sensitivity analysis tends to zero or one as $\Gamma > \tilde{\Gamma}$ or $\Gamma < \tilde{\Gamma}$ where $\tilde{\Gamma}$ is the design sensitivity.

Design Sensitivity $\tilde{\Gamma}$ For Instruments with Varying Strength. Effect size $\lambda = (\beta_0 - \beta) / \sigma$

Compliance		100%	50%	20%	10%
π_A, π_C, π_N		0, 1, 0	$\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$	$\frac{2}{5}, \frac{1}{5}, \frac{2}{5}$	$\frac{9}{20}, \frac{2}{20}, \frac{9}{20}$
ϵ_i	λ				
Normal	1	11.7	2.7	1.5	1.2
Normal	$\frac{1}{2}$	3.2	1.7	1.2	1.1
Cauchy	1	3.0	1.7	1.2	1.1
Cauchy	$\frac{1}{2}$	1.8	1.4	1.1	1.1
Logistic	1	3.9	1.9	1.3	1.1
Logistic	$\frac{1}{2}$	2.0	1.4	1.1	1.1

- Large effect, perfect compliance, Normal errors: $(\beta_0 - \beta) / \sigma = 1$, $(\pi_A, \pi_C, \pi_N) = (0, 1, 0)$ yields $\tilde{\Gamma} = 11.7$.
- In table, $\tilde{\Gamma} \geq 1.4$ for a strong instrument: $(\pi_A, \pi_C, \pi_N) = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$.
- In table, weak instruments are always sensitive to unobserved biases, $(\pi_A, \pi_C, \pi_N) = (\frac{9}{20}, \frac{2}{20}, \frac{9}{20})$.

53 Interpretation of IV Design Sensitivity Results

- Somewhat contextual. Strength of an instrument and plausible biases depend upon context.
- Weak instruments are always sensitive to small biases. They may also carry limited information even without bias.
- Strong instruments may or may not be insensitive to small biases.
- A strong instrument may be preferable even when small biases are possible.
- Comparing men born a few years apart.

54 Summary

- Literature on observational studies contains informal advice, often good advice, about designs to strengthen causal conclusions.
- Issues like dose-response, coherence, reducing heterogeneity, instrumental variables of various types.
- It is possible to formalize these considerations in terms of design sensitivity. What aspects of design are expected to reduce sensitivity to unobserved biases when a sensitivity analysis is preformed.
- Quantitative rather than qualitative.
- Greater precision about what advice says and when it is likely to work.

55 A0: Specific Results

Heterogeneity: An important issue, strongly affecting design sensitivity.

Doses: Dose-response has a small effect compared to uniform at the average dose. Given the choice, it would be better to make the doses uniform but larger.

Coherence: In principle, could have a substantial impact, but its importance depends strongly on inter-correlation.

IV: Weak instruments are always sensitive to small biases. Strong instruments may resist small to moderate biases. Strict validity may be overrated compared to strength.

56 A1: 2 Years of Age and Earnings

- The 1926-1928 pairs differed by 2 year of age and 76% of the 1926 men served in WWII, while 24% of the 1928 men served. The IV argument attributes the difference in earnings to the difference in service. Could it be the difference in age, not the difference in earnings?
- The 1926-1924 pairs differed by 2 year of age and 76% of the 1926 men served in WWII, while 78% of the 1924 men served. So the difference in service is small, but the difference in age is the same.
- Using Wilcoxon's signed rank statistic again, the 95% confidence interval for the typical difference in annual earnings, 1926-1924, was $[-250, 240]$. So the change in age with no change in service was associated with a very small, possible zero, shift in earnings.

57 A2: Practical Illustrations

- In practice, can't know for certain about unobserved biases, but can use tactics that are likely to reduce heterogeneity, perhaps at the expense of sample size.
- Tactics that attempt to reduce unobserved bias may reduce heterogeneity.
- In both cases, we are trying to arrange things to compare units that are similar in relevant ways we have not observed.
- Can recognize and employ tactics aimed at this goal, but can't be certain whether they reduced unobserved bias, heterogeneity, both or neither.

58 A3: Returns to Education

- Economic returns to additional education.
- Can't just compare high school dropouts and college graduates. They differed in terms of parents wealth and education, possibly genetic endowment.
- Would like to compare children of the same parents, growing up at the same time in the same home with the same genes.
- Ashenfelter & Rouse (1998) compared identical twins with differing educations, estimating a 9% increase in earnings per year of additional education.

59 A4: Road Hazards

- What permanent road hazards increase risk of fatal collisions with roadside objects? Road hazards are a small part of the total picture. Also important:

Driver: Driver's skill, aggressiveness, risk tolerance, sobriety.

Weather: Ice, snow, rain, fog, ambient light.

Safety equipment: Brakes, tires, traction control, stability control, air bags, use of seat belts.

Related: Sobriety more common at noon than midnight, so sobriety and ambient light related. In rain or snow, drive on highway to work, but not on dirt road to picnic area or hiking trail, so weather and roadside hazards vary together.

60 A5: Road Hazards: a case-crossover study

- Would like to compare different road hazards with the same driver, in the same state of sobriety, in the same car, in the same weather, with the same ambient light, with seat belts in the same state of use. Is this possible?
- Wright and Robertson (1976) examined 300 fatal accidents involving a collision with a roadside object (trees, embankments, ditches, etc.) in Georgia 1974-1975.
- Compared these to 300 non-accidents involving the same driver, car, weather, light, etc. There were 1 mile back along the road, a location passed by the driver minutes before the crash.
- Crash sites had a substantial excess of roads curving more than six degrees with downhill gradients greater than 2%.

61 A6: Minimum Wage Laws

- Do minimum wage laws reduce employment?
- Traditional to study this using states and/or time-periods with different minimum wage laws. But businesses vary between states, and business conditions vary with time.
- Would like to compare nearly identical businesses in states with different minimum wage laws. How does one find nearly identical businesses?
- Card and Krueger (1994) looked at changes, after-minus-before, in employment in NJ and PA when NJ increased its minimum wage by 19% in 1992. They looked at fast food restaurants, comparing Burger Kings to Burger Kings, Wendy's to Wendy's, etc. Found no sign of reduced employment..

62 A7: Motorcycle helmets

- To what extent do helmets reduce risk of death in motorcycle crashes?
- Crashes vary: speeds, forces, traffic density, other vehicles, etc.
- Would like to compare two people, on the same type of motorcycle, riding at the same speed, on the same road, in the same traffic, crashing into the same object. Is this possible?
- It is when two people ride the same motorcycle, one with, the other without a helmet. Norvell and Cummings (2002) looked at such crashes, estimating a 40% reduction in fatality risk associated with helmet use.