Heterogeneity and Causality: Unit Heterogeneity and Design Sensitivity in Observational Studies

Paul R. Rosenbaum, University of Pennsylvania

References

[1] Rosenbaum, P. R. (2005) Heterogeneity and causality: Unit heterogeneity and design sensitivity in observational studies. *American Statistician*, 59, 147-152.

[2] Rosenbaum, P. R. (2004) Design sensitivity in observational studies. *Biometrika*, 91, 153-164.

[3] Mill, J. S. (1867), *A System of Logic: The Principles of Evidence and the Methods of Scientific Investigation*, New York: Harper & Brothers.

[4] Fisher, R. A. (1935), *Design of Experiments*, Edinburgh: Oliver and Boyd.

1 Some terms

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

Unit heterogeneity: More or less, for units with the same observed covariates, the dispersion of the responses units would exhibit if the units received the control.

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, treated and control groups may look similar in terms of observed covariates, but may differ in terms of 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 Outline of Talk

Introduction: Mostly motivation.

Review: Notation, role of randomization in experiments, sensitivity analysis in observational studies.

Theoretical point demonstrated: Done three ways: (i) one simulated data set, (ii) statistical power of competing designs, (iii) theorem about the limiting case.

Practical point illustrated: Some actual studies which seem to have done a good job with issue.

3 A Distinction

LM versus SL

- An easy distinction to make.

- Some theory might be interpreted to suggest the distinction is not extremely interesting.

- In the simplest, stylized illustrations of this distinction (Gaussian distributions with errors having known constant variance), the mle's have the same distributions, as do confidence intervals and significance levels; hence also power.

- So the distinction is starting to look like a yawn.

- And yet, the distinction is enormously important.

4 A Distinction, continued

LM versus SL

- The little bit of theory, that might be interpreted as suggesting the yawn, more or less assumes a randomized experiment, more or less assumes no unobserved biases, no biases of $O(1)$. (Why 'more or less'? The assumption is not explicit; rather, the possibility that the assumption isn't true isn't mentioned.)

- And yet, LM and SL differ dramatically in their sensitivity to unobserved biases. SL is much better.

- Worse yet, unaided by statistical theory, the gut instinct is to go for LM.

5 Simplest Situation: Matched Pairs

- I pairs, $i = 1, \ldots, I$, of two units, $j = 1, 2$, one treated, one control, matched for observed covariates, yielding I treated-minus-control differences in responses, D_i, $i = 1, \ldots, I$.

- In a randomized experiment, there is a sense (formalized soon) that the D_i estimate the effect of the treatment. Also, they form the basis for randomization inferences about treatment effect in Fisher's (1935) sense.

- In an observational study, units might, for instance, decide for themselves whether to take the treatment. Because of this, the D_i's might tend to be positive even if the treatment had no effect.

- In one example later, the pairs are people with differing education, and D_i measures the difference in their earnings. Worry that people who get more education are different.

6 Simplest Situation, with Good Fortune

- Imagine that we were very lucky: our observational study was free of unobserved biases, and controlling for observed covariates was enough to estimate treatment effect, specifically a positive constant effect, say $\tau > 0$.

- We would not know this from data. Many D_i's would be positive, but that could be a positive effect or a positive bias.

- The best we could hope for would be that the sensitivity analysis would report back that the results were insensitive to small and moderate biases.

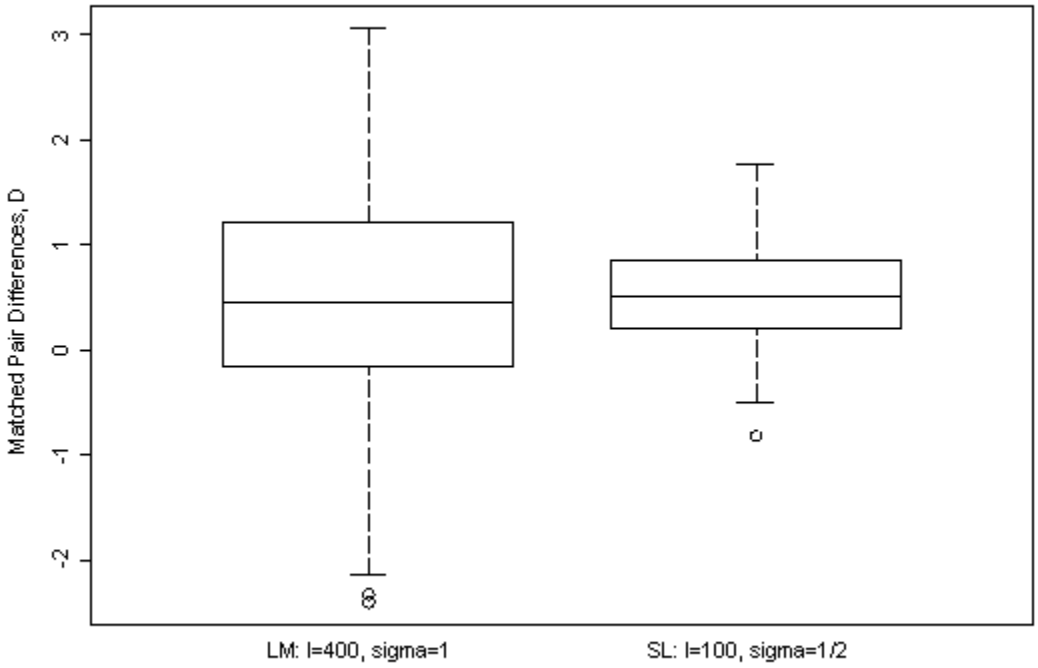
7 What is the distinction?

- LM (larger sample size, more heterogeneous): $D_i \underset{iid}{\sim} N\left(\tau, \sigma^2\right)$ with sample size $4I$ pairs.

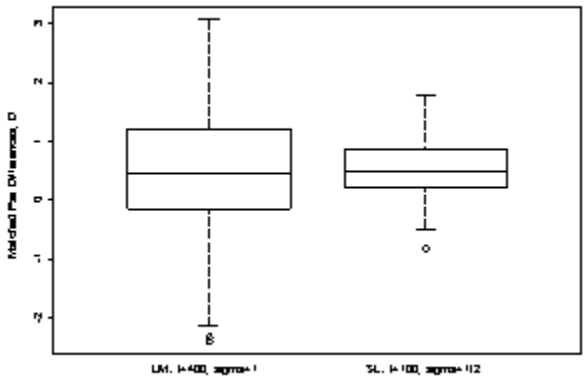
- SL (smaller sample size, less heterogeneous): $D_i \underset{iid}{\sim} N\left(\tau, \eta^2\right)$ with sample size I pairs with $\eta = \sigma/2$.

- Then the sample mean is $\overline{D} \sim N\left(\tau, \sigma^2/4I\right)$ in both LM and SL.

- If we were not worried about unobserved biases, the distinction would not matter much.



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.



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 $\widehat{\tau}$	0.50	0.52
Wilcoxon 95% CI $[\widehat{\tau}_L,\ \widehat{\tau}_H]$	[.40, .60]	[.43, .62]

8 Notation 1

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

Matching: I pairs, $i = 1, \ldots, I$, 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 $i = 1, \ldots, I$.

Responses: Two potential responses, $\left(r_{Tij}, r_{Cij}\right)$, r_{Tij} under treatment, $Z_{ij} = 1$, r_{Cij} under control, $Z_{ij} = 0$, so effect is $r_{Tij} - r_{Cij}$; Neyman (1935) and Rubin (1974).

9 Paired Randomized Experiment

Conditioning: Write

$$\begin{aligned}\mathcal{F} &= \left\{\left(r_{Tij}, r_{Cij}, \mathbf{x}_{ij}, u_{ij}\right), i = 1, \ldots, I, j = 1, 2\right\} \\ \mathcal{Z} &= \left\{Z_{i1} + Z_{i2} = 1,\ i = 1, \ldots, I\right\};\end{aligned}$$

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

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

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

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 $\epsilon_i = (2Z_{i1} - 1)\,(r_{Ci1} - r_{Ci2})$.

10 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 $\frac{1}{2}$, $i = 1, \ldots, I$.

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 $\widehat{\tau}$ such that $W_{\widehat{\tau}}$ is as close as possible to its null expectation, $I\ (I+1)/4$.

11 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: $(r_{Ci1} - r_{Ci2})/\sigma \sim_{iid} F(\cdot)$ were $\sigma > 0$ 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).

12 Departures from Random Assignment

- 1. In the population prior to matching, treatment assignments were independent, with unknown probabilities $\pi_{ij} = \mathsf{Pr}\left(Z_{ij} = 1 \,\middle|\, \mathcal{F}\right)$

  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 $\Gamma \geq 1$,

$$\frac{1}{\Gamma} \leq \frac{\pi_{ij}\left(1-\pi_{ik}\right)}{\pi_{ik}\left(1-\pi_{ij}\right)} \leq \Gamma, \;\; \forall\, i, j, k \qquad (1)$$

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

13 Departures, continued

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

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

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

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 $\Pr\left(Z_{i1} = 1 \mid \mathcal{Z}, \mathcal{F}\right)$ 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?

14 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 $1-\theta$, $i = 1, \ldots, I$; 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

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

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

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

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

15 Procedure, continued

For each $\Gamma \geq 1$

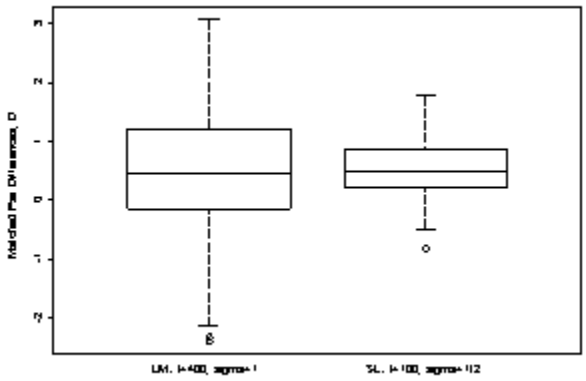
$$\Pr\left(\overline{W} \geq w\right) \leq \Pr\left(W_{\tau_0} \geq w \mid \mathcal{Z}, \mathcal{F}\right) \leq \Pr\left(\overline{\overline{W}} \geq w\right)$$

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 $W_{\tau_0} \geq \widetilde{w}$ where $0.05 = \Pr\left(\overline{\overline{W}} \geq \widetilde{w}\right)$.

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

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

yield an interval of HL point estimates, $[\widehat{\tau}_{\mathsf{min}}, \widehat{\tau}_{\mathsf{max}}]$. With no unobserved bias, $\Gamma = 1$, $\mu_{\boldsymbol{\pi}} = I\left(I+1\right)/4$, and $\widehat{\tau}_{\mathsf{min}} = \widehat{\tau}_{\mathsf{max}}$ 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 $\widehat{\tau}$	0.50	0.52
Wilcoxon 95% CI $[\widehat{\tau}_L,\ \widehat{\tau}_H]$	[.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
$\Gamma = 1$	$\widehat{\tau} = \widehat{\tau}_{\text{min}} = \widehat{\tau}_{\text{max}}$	0.50	0.52
$\Gamma = 2$	$[\widehat{\tau}_{\text{min}},\ \widehat{\tau}_{\text{max}}]$	[.19, .81]	[.37, .67]

16 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

17 Limiting case, $I \to \infty$

- As the number of pairs $I \to \infty$, the only uncertainty that remains is due to unobserved bias.

- In particular, for each $\Gamma \geq 1$, as $I \to \infty$, the (random) interval of HL estimates, $[\widehat{\tau}_{\mathsf{min}}, \widehat{\tau}_{\mathsf{max}}]$ converges in probability to a fixed interval $[\tau_{\mathsf{min}}, \tau_{\mathsf{max}}]$.

- If $\Gamma = 1$, then $\tau_{\mathsf{min}} = \tau_{\mathsf{max}} = \tau$.

- If $\Gamma > 1$, then $\tau_{\mathsf{min}} < \tau_{\mathsf{max}}$, with $\tau \in [\tau_{\mathsf{min}}, \tau_{\mathsf{max}}]$.

18 Limiting case, $I \to \infty$

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 $[\tau_{\text{min}},\ \tau_{\text{max}}]$?

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

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

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

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

19 Theoretical Point

LM:

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

SL:

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

In a randomized experiment: Not a big difference.

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

20 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.

21 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.

22 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.

23 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%.

24 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..

25 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.

26 Summary

- If treatments are randomly assigned, unbiased estimates of effects are available. Increasing sample size and reducing unit heterogeneity reduce standard errors of unbiased estimates.

- Without randomization, unobserved biases are possible, perhaps likely. Reducing heterogeneity in responses, even purely random heterogeneity, confers benefits that cannot be had by increasing the sample size.

- Specifically, reducing heterogeneity reduces sensitivity to unobserved bias.

- Examples illustrated tactics that have been used that are likely to reduce heterogeneity.