Being Realistic About Unmeasured Biases in Observational Studies

Paul R. Rosenbaum

January 2026

Terminology

- **Definition of an Observational Study**: A study of the effects caused by competing treatments that were not randomly assigned to individuals (Cochran 1965, JRSS-A, “Planning of observational studies of human populations”).

Terminology

- **Definition of an Observational Study**: A study of the effects caused by competing treatments that were not randomly assigned to individuals (Cochran 1965, JRSS-A, “Planning of observational studies of human populations”).
- Treatment and outcome may be **associated in the absence of an effect caused by the treatment**, because treatments were not randomly assigned.

Terminology

- **Definition of an Observational Study**: A study of the effects caused by competing treatments that were not randomly assigned to individuals (Cochran 1965, JRSS-A, “Planning of observational studies of human populations”).
- Treatment and outcome may be **associated in the absence of an effect caused by the treatment**, because treatments were not randomly assigned.
- Although we always adjust for measured covariates, treated and control groups may nonetheless **differ in terms of covariates that were not measured**.

Terminology

- **Definition of an Observational Study**: A study of the effects caused by competing treatments that were not randomly assigned to individuals (Cochran 1965, JRSS-A, “Planning of observational studies of human populations”).
- Treatment and outcome may be **associated in the absence of an effect caused by the treatment**, because treatments were not randomly assigned.
- Although we always adjust for measured covariates, treated and control groups may nonetheless **differ in terms of covariates that were not measured**.
- **That is**: without random assignment, the probability of treatment may depend upon relevant covariates that were not measured.

Terminology

- **Definition of an Observational Study**: A study of the effects caused by competing treatments that were not randomly assigned to individuals (Cochran 1965, JRSS-A, “Planning of observational studies of human populations”).
- Treatment and outcome may be **associated in the absence of an effect caused by the treatment**, because treatments were not randomly assigned.
- Although we always adjust for measured covariates, treated and control groups may nonetheless **differ in terms of covariates that were not measured**.
- **That is**: without random assignment, the probability of treatment may depend upon relevant covariates that were not measured.
- This is **the main source of controversy** in observational studies, and it **organizes the design and analysis** of an observational study.

An example, some methods, some theory for design and analysis

- In thinking about unmeasured biases, **context matters**. The talk uses a simple example with **a context that is familiar to everyone**.

An example, some methods, some theory for design and analysis

- In thinking about unmeasured biases, **context matters**. The talk uses a simple example with **a context that is familiar to everyone**.
- Some simple quick claims about how observational studies should be **designed** if they are **to have greater insensitivity to unmeasured biases**. (Proofs of these claims are in Parts III and IV of my *Design of Observational Studies*, 2^{nd} edition, 2020.)

An example, some methods, some theory for design and analysis

- In thinking about unmeasured biases, **context matters**. The talk uses a simple example with **a context that is familiar to everyone**.
- Some simple quick claims about how observational studies should be **designed** if they are **to have greater insensitivity to unmeasured biases**. (Proofs of these claims are in Parts III and IV of my *Design of Observational Studies*, 2^{nd} edition, 2020.)
- Some theory showing that **choice of methods of analysis** has a substantial effect on the degree to which a study is sensitive to unmeasured biases.

An example, some methods, some theory for design and analysis

- In thinking about unmeasured biases, **context matters**. The talk uses a simple example with **a context that is familiar to everyone**.
- Some simple quick claims about how observational studies should be **designed** if they are **to have greater insensitivity to unmeasured biases**. (Proofs of these claims are in Parts III and IV of my *Design of Observational Studies*, 2^{nd} edition, 2020.)
- Some theory showing that **choice of methods of analysis** has a substantial effect on the degree to which a study is sensitive to unmeasured biases.
- Perhaps surprisingly, **evidence of unmeasured bias** may make an observational study insensitive to larger unmeasured biases.

An example, some methods, some theory for design and analysis

- In thinking about unmeasured biases, **context matters**. The talk uses a simple example with **a context that is familiar to everyone**.
- Some simple quick claims about how observational studies should be **designed** if they are **to have greater insensitivity to unmeasured biases**. (Proofs of these claims are in Parts III and IV of my *Design of Observational Studies*, 2^{nd} edition, 2020.)
- Some theory showing that **choice of methods of analysis** has a substantial effect on the degree to which a study is sensitive to unmeasured biases.
- Perhaps surprisingly, **evidence of unmeasured bias** may make an observational study insensitive to larger unmeasured biases.
- The example has **several control groups**, so the logic of several control groups will be briefly discussed.

Derived from:

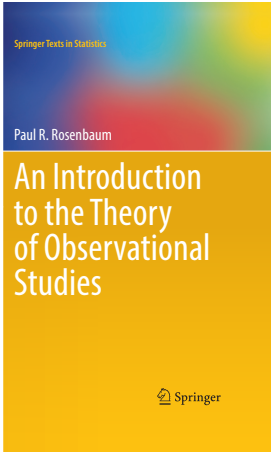

Figure: Data in `R` package `iTOS`.

Some General Theses

- **Every observational study is affected by unmeasured biases**, but that fact is not debilitating. Example: smoking and lung cancer.

Some General Theses

- **Every observational study is affected by unmeasured biases**, but that fact is not debilitating. Example: smoking and lung cancer.
- Unmeasured bias is unmeasured, but it often has detectable consequences. The **detectable consequences may heighten or diminish concern** that the ostensible causal effects are spurious.

Some General Theses

- **Every observational study is affected by unmeasured biases**, but that fact is not debilitating. Example: smoking and lung cancer.
- Unmeasured bias is unmeasured, but it often has detectable consequences. The **detectable consequences may heighten or diminish concern** that the ostensible causal effects are spurious.
- A **sensitivity analysis talks** about unmeasured biases, but it is computed from – it **is a function of – observable data** from observable distributions. Change the observable distributions – change the study design – change the analysis and you change the sensitivity to unmeasured biases.

Some General Theses

- **Every observational study is affected by unmeasured biases**, but that fact is not debilitating. Example: smoking and lung cancer.
- Unmeasured bias is unmeasured, but it often has detectable consequences. The **detectable consequences may heighten or diminish concern** that the ostensible causal effects are spurious.
- A **sensitivity analysis talks** about unmeasured biases, but it is computed from – it **is a function of – observable data** from observable distributions. Change the observable distributions – change the study design – change the analysis and you change the sensitivity to unmeasured biases.
- **Without guidance from statistical theory** about the previous point, it is easy to make poor decisions in design and analysis, reporting that your results are sensitive to small biases when they are not.

Example: HDL Cholesterol and Light Daily Alcohol

- You often hear or read that a glass of wine each day with dinner prolongs life reducing cardiovascular mortality, perhaps by increasing HDL cholesterol levels (e.g., Suh et al. *Ann. Int. Med.* 1992;116:881-887.

Example: HDL Cholesterol and Light Daily Alcohol

- You often hear or read that a glass of wine each day with dinner prolongs life reducing cardiovascular mortality, perhaps by increasing HDL cholesterol levels (e.g., Suh et al. *Ann. Int. Med.* 1992;116:881-887.
- A recent position paper by the American Society of Clinical Oncology (Noel Loconte et al. *J. Clin. Oncol.* 2018;36:83-93) is sharply critical of this claim, emphasizing increased risk of death from cancer, although risks from accidents, liver diseases, and violence are relevant too.

Example: HDL Cholesterol and Light Daily Alcohol

- You often hear or read that a glass of wine each day with dinner prolongs life reducing cardiovascular mortality, perhaps by increasing HDL cholesterol levels (e.g., Suh et al. *Ann. Int. Med.* 1992;116:881-887.
- A recent position paper by the American Society of Clinical Oncology (Noel Loconte et al. *J. Clin. Oncol.* 2018;36:83-93) is sharply critical of this claim, emphasizing increased risk of death from cancer, although risks from accidents, liver diseases, and violence are relevant too.
- Purely as a methodological example, will look at a small corner (and alas less important) corner of this topic, namely whether light daily alcohol consumption increases HDL cholesterol.

Example: HDL Cholesterol and Light Daily Alcohol

- You often hear or read that a glass of wine each day with dinner prolongs life reducing cardiovascular mortality, perhaps by increasing HDL cholesterol levels (e.g., Suh et al. *Ann. Int. Med.* 1992;116:881-887.
- A recent position paper by the American Society of Clinical Oncology (Noel Loconte et al. *J. Clin. Oncol.* 2018;36:83-93) is sharply critical of this claim, emphasizing increased risk of death from cancer, although risks from accidents, liver diseases, and violence are relevant too.
- Purely as a methodological example, will look at a small corner (and alas less important) corner of this topic, namely whether light daily alcohol consumption increases HDL cholesterol.
- For some discussion of mortality and light alcohol consumption, see my: "Does a daily glass of wine prolong life? Insight from a second control group," *Chance*, 2025;38(1):25-30.

Example: Treated and Control Groups

- Adults, age ≥ 20, from NHANES 2013-2016. (At this time, NHANES and CDC defined "binge drinking" as ≥ 4 or 5 drinks in a day $\doteq$ legally drunk.)

Example: Treated and Control Groups

- Adults, age ≥ 20, from NHANES 2013-2016. (At this time, NHANES and CDC defined "binge drinking" as ≥ 4 or 5 drinks in a day $\doteq$ legally drunk.)
- Treated group consumed between 1 and 3 alcoholic drinks on most days, meaning on $\geq 260 = 5 \times 52$ days last year. (median 520 drinks/year)

Example: Treated and Control Groups

- Adults, age ≥ 20, from NHANES 2013-2016. (At this time, NHANES and CDC defined "binge drinking" as ≥ 4 or 5 drinks in a day $\doteq$ legally drunk.)
- Treated group consumed between 1 and 3 alcoholic drinks on most days, meaning on $\geq 260 = 5 \times 52$ days last year. (median 520 drinks/year)
- Control group N (=Never) had fewer than 12 drinks in their life. (median 0/year).

Example: Treated and Control Groups

- Adults, age ≥ 20, from NHANES 2013-2016. (At this time, NHANES and CDC defined "binge drinking" as ≥ 4 or 5 drinks in a day $\doteq$ legally drunk.)
- Treated group consumed between 1 and 3 alcoholic drinks on most days, meaning on $\geq 260 = 5 \times 52$ days last year. (median 520 drinks/year)
- Control group N (=Never) had fewer than 12 drinks in their life. (median 0/year).
- Control group R (=Rarely) had more than 12 drinks in their life, but fewer than 12 drinks in the past year. Never had a period in their lives when they engaged in binge drinking on most days. (median 0 drinks/year).

Example: Treated and Control Groups

- Adults, age ≥ 20, from NHANES 2013-2016. (At this time, NHANES and CDC defined "binge drinking" as ≥ 4 or 5 drinks in a day $\doteq$ legally drunk.)
- Treated group consumed between 1 and 3 alcoholic drinks on most days, meaning on $\geq 260 = 5 \times 52$ days last year. (median 520 drinks/year)
- Control group N (=Never) had fewer than 12 drinks in their life. (median 0/year).
- Control group R (=Rarely) had more than 12 drinks in their life, but fewer than 12 drinks in the past year. Never had a period in their lives when they engaged in binge drinking on most days. (median 0 drinks/year).
- Control group B (=former Binge drinker) had a period in their lives when they engaged in binge drinking on most days, but stopped, and currently drinks, if at all, on at most one day a week (i.e., 52 days in the past year). (median 4 drinks/year)

Example: Treated and Control Groups

- Adults, age ≥ 20, from NHANES 2013-2016. (At this time, NHANES and CDC defined "binge drinking" as ≥ 4 or 5 drinks in a day $\doteq$ legally drunk.)
- Treated group consumed between 1 and 3 alcoholic drinks on most days, meaning on $\geq 260 = 5 \times 52$ days last year. (median 520 drinks/year)
- Control group N (=Never) had fewer than 12 drinks in their life. (median 0/year).
- Control group R (=Rarely) had more than 12 drinks in their life, but fewer than 12 drinks in the past year. Never had a period in their lives when they engaged in binge drinking on most days. (median 0 drinks/year).
- Control group B (=former Binge drinker) had a period in their lives when they engaged in binge drinking on most days, but stopped, and currently drinks, if at all, on at most one day a week (i.e., 52 days in the past year). (median 4 drinks/year)
- Take a moment and think about people in these groups.

Observational Block Design

- $I = 406$ blocks of size $J = 4$, one person from each group, matched for age, sex, and education (1 is $<$9th grade, 3 is high school, 5 is $\geq$ BA degree.),

Observational Block Design

- $I = 406$ blocks of size $J = 4$, one person from each group, matched for age, sex, and education (1 is $<$9th grade, 3 is high school, 5 is $\geq$ BA degree.),
- **Plus** a binary indicator of whether they were in a NHANES subsample that measured methylmercury levels in blood (200 blocks yes, 206 blocks no).

Table: Covariates **Before=Be** and **After=Af** matching, and the remainder that was **Not** matched . D=daily, N=never, R=rarely, B=past binger. All D's were matched.

	Sample Size			Female %			Age			Education		
	Be	Af	Not	Be	Af	Not	Be	Af	Not	Be	Af	Not
D	406	406	0	**34**	34		57	57		**4.1**	4.1	
N	1536	406	1130	**71**	34	84	51	57	50	3.2	3.8	2.9
R	1237	406	831	**72**	34	90	53	56	51	3.4	3.9	3.2
B	914	406	508	**29**	34	25	54	56	53	3.1	3.9	**2.5**

Brief Mention of Design Techniques to Address Unmeasured Biases

- Campbell/Bitterman idea that **multiple control groups** cannot control unmeasured biases, but they can systematically vary them to see if they matter.

Brief Mention of Design Techniques to Address Unmeasured Biases

- Campbell/Bitterman idea that **multiple control groups** cannot control unmeasured biases, but they can systematically vary them to see if they matter.
- Comparison of daily drinkers with people who barely drink, omitting people who drink twice a week. **Omitting diluted versions of the treatment** increases insensitivity to unmeasured biases (*Design of Observational Studies*, 2020, §18.4; *Introduction to the Theory of Observational Studies* (*iTOS*), 2025, §10.3)

Brief Mention of Design Techniques to Address Unmeasured Biases

- Campbell/Bitterman idea that **multiple control groups** cannot control unmeasured biases, but they can systematically vary them to see if they matter.
- Comparison of daily drinkers with people who barely drink, omitting people who drink twice a week. **Omitting diluted versions of the treatment** increases insensitivity to unmeasured biases (*Design of Observational Studies*, 2020, §18.4; *Introduction to the Theory of Observational Studies* (*iTOS*), 2025, §10.3)
- **Blocks of size 4 are a better design** (1-treated-to-3-controls), better for example than pairs (even many more pairs). Selection bias is harder to distinguish from a treatment effect in pairs or unmatched comparisons, and easier to distinguish with 1-to-3 blocks. (JASA 2024, Biometrics 2013;69:118-127, *iTOS*, 2025, §10.2).

Brief Mention of Design Techniques to Address Unmeasured Biases

- Campbell/Bitterman idea that **multiple control groups** cannot control unmeasured biases, but they can systematically vary them to see if they matter.
- Comparison of daily drinkers with people who barely drink, omitting people who drink twice a week. **Omitting diluted versions of the treatment** increases insensitivity to unmeasured biases (*Design of Observational Studies*, 2020, §18.4; *Introduction to the Theory of Observational Studies* (*iTOS*), 2025, §10.3)
- **Blocks of size 4 are a better design** (1-treated-to-3-controls), better for example than pairs (even many more pairs). Selection bias is harder to distinguish from a treatment effect in pairs or unmatched comparisons, and easier to distinguish with 1-to-3 blocks. (JASA 2024, Biometrics 2013;69:118-127, *iTOS*, 2025, §10.2).
- **An unaffected outcome, methylmercury**. WHO & CDC say almost all human exposure to methylmercury comes from eating fish/shellfish. Those who have looked for methylmercury in alcoholic beverages haven't found it. Can we use this?

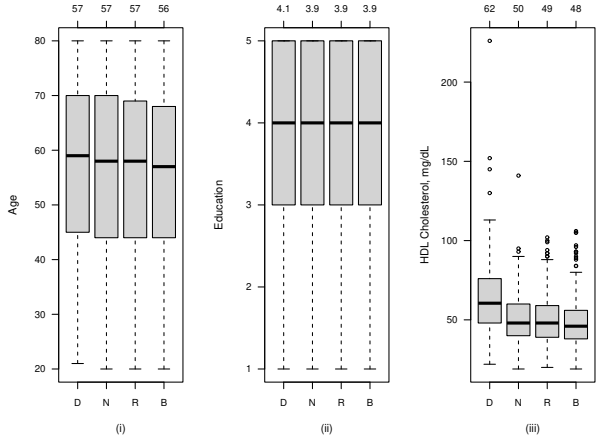

Figure: $I = 406$ matched blocks. Each group is 33.7% female. M-estimates of location are at the top. D = daily drinking, N = never, R = rare, B = formerly a frequent binge drinker. 6 Pairwise Holm comparisons: D-vs-each control, $P \leq 10^{-16}$, each control-vs-control, $P \geq 0.21$.

Do you think the groups are living similar lives?

- First thesis was: In observational studies, there are always unmeasured biases.

Do you think the groups are living similar lives?

- First thesis was: In observational studies, there are always unmeasured biases.
- Tests use Friedman or Cochran Q

Table: Blocked comparisons. $\overline{X}$ is the mean, M is the median.

Variable		Alcohol Group				
D=daily, N=never, R=rarely, B=past binge		D	N	R	B	P-value
Ever tried marijuana or hashish?	%	73	9	25	75	0.0000000
Ever tried cocaine, heroin, meth?	%	29	4	4	37	0.0000000
Methylmercury in blood (μg/L)	M	1.12	0.54	0.56	0.56	0.0000008
Been to dentist in past year?	%	67	58	57	48	0.0000006

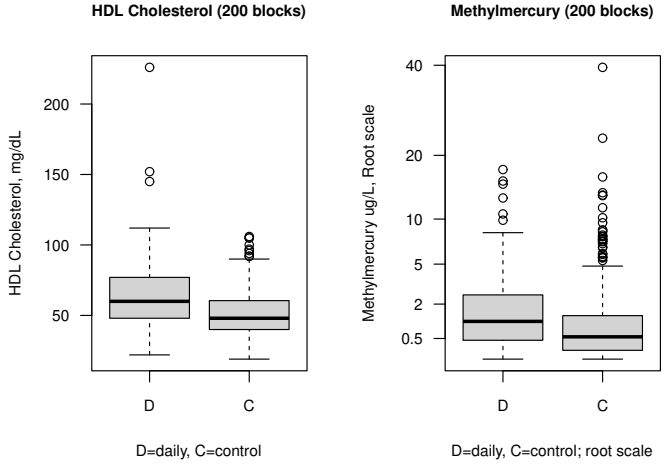

Figure: 200 blocks with methylmercury data. $\sqrt{y}$ scale on right. Control groups are merged.

Basic Structure: Treatments, Covariates, Outcomes

- **Treatments**: Treated if $Z = 1$ or control if $Z = 0$.

Basic Structure: Treatments, Covariates, Outcomes

- **Treatments**: Treated if $Z = 1$ or control if $Z = 0$.
- **Causal effects**: (Neyman 1923, Rubin 1974) Comparison of a **potential** outcome r_T under treatment, seen if $Z = 1$, and a potential outcome under control, r_C, seen if $Z = 0$, so we observe from a person $(R,\ Z)$ for a person, where $R = Z\, r_T + (1 - Z)\, r_C$.

Basic Structure: Treatments, Covariates, Outcomes

- **Treatments**: Treated if $Z = 1$ or control if $Z = 0$.
- **Causal effects**: (Neyman 1923, Rubin 1974) Comparison of a **potential** outcome r_T under treatment, seen if $Z = 1$, and a potential outcome under control, r_C, seen if $Z = 0$, so we observe from a person $(R,\ Z)$ for a person, where $R = Z\, r_T + (1 - Z)\, r_C$.
- Outcomes r_T, r_C and R may be **multivariate**. (HDL cholesterol, methymercury).

Basic Structure: Treatments, Covariates, Outcomes

- **Treatments**: Treated if $Z = 1$ or control if $Z = 0$.
- **Causal effects**: (Neyman 1923, Rubin 1974) Comparison of a **potential** outcome r_T under treatment, seen if $Z = 1$, and a potential outcome under control, r_C, seen if $Z = 0$, so we observe from a person $(R,\ Z)$ for a person, where $R = Z\, r_T + (1 - Z)\, r_C$.
- Outcomes r_T, r_C and R may be **multivariate**. (HDL cholesterol, methymercury).
- **Covariates**: We also observe a covariate $\mathbf{x}$ and are concerned about unobserved covariates u.

Basic Structure: Treatments, Covariates, Outcomes

- **Treatments**: Treated if $Z = 1$ or control if $Z = 0$.
- **Causal effects**: (Neyman 1923, Rubin 1974) Comparison of a **potential** outcome r_T under treatment, seen if $Z = 1$, and a potential outcome under control, r_C, seen if $Z = 0$, so we observe from a person $(R,\ Z)$ for a person, where $R = Z\, r_T + (1 - Z)\, r_C$.
- Outcomes r_T, r_C and R may be **multivariate**. (HDL cholesterol, methymercury).
- **Covariates**: We also observe a covariate $\mathbf{x}$ and are concerned about unobserved covariates u.
- **Randomized experiment**: Z is determined by a coin flip, perhaps after blocking or matching for some function $\mathbf{h}(\mathbf{x})$. The coin is "fair" in not depending upon $(r_T,\ r_C)$, or more precisely ...

Ignorable Treatment Assignment & Principal Unobserved Covariate

- Treatment assignment is **ignorable given the observed covariates x** if

$$0 < \Pr(Z = 1 \mid \mathbf{x},\, r_T,\, r_C) = \Pr(Z = 1 \mid \mathbf{x}) < 1 \tag{1}$$

Ignorable Treatment Assignment & Principal Unobserved Covariate

- Treatment assignment is **ignorable given the observed covariates $\mathbf{x}$** if

$$0 < \Pr(Z = 1 \mid \mathbf{x},\, r_T,\, r_C) = \Pr(Z = 1 \mid \mathbf{x}) < 1 \tag{1}$$

- If ignorable, **adjustments** for $\mathbf{x}$ **suffice** for causal inference.

Ignorable Treatment Assignment & Principal Unobserved Covariate

- Treatment assignment is **ignorable given the observed covariates $\mathbf{x}$** if

$$0 < \Pr(Z = 1 \mid \mathbf{x},\, r_T,\, r_C) = \Pr(Z = 1 \mid \mathbf{x}) < 1 \tag{1}$$

- If ignorable, **adjustments** for $\mathbf{x}$ **suffice** for causal inference.
- We often speak of **ignorable assignment given something else**, given a function $\mathbf{h}(\mathbf{x})$ of $\mathbf{x}$, or given $(\mathbf{x},\, u)$ where u is an unobserved covariate.

Ignorable Treatment Assignment & Principal Unobserved Covariate

- Treatment assignment is **ignorable given the observed covariates $\mathbf{x}$** if

$$0 < \Pr(Z = 1 \mid \mathbf{x},\, r_T,\, r_C) = \Pr(Z = 1 \mid \mathbf{x}) < 1 \qquad (1)$$

- If ignorable, **adjustments** for $\mathbf{x}$ **suffice** for causal inference.
- We often speak of **ignorable assignment given something else**, given a function $\mathbf{h}(\mathbf{x})$ of $\mathbf{x}$, or given $(\mathbf{x},\, u)$ where u is an unobserved covariate.
- $\Pr(Z = 1 \mid \mathbf{x}) = e(\mathbf{x})$, say, is the **propensity score**, and (1) $\Longrightarrow$ ignorable given $e(\mathbf{x})$.

Ignorable Treatment Assignment & Principal Unobserved Covariate

- Treatment assignment is **ignorable given the observed covariates $\mathbf{x}$** if

$$0 < \Pr(Z = 1 \,|\, \mathbf{x},\, r_T,\, r_C) = \Pr(Z = 1 \,|\, \mathbf{x}) < 1 \qquad (1)$$

- If ignorable, **adjustments** for $\mathbf{x}$ **suffice** for causal inference.
- We often speak of **ignorable assignment given something else**, given a function $\mathbf{h}(\mathbf{x})$ of $\mathbf{x}$, or given $(\mathbf{x},\, u)$ where u is an unobserved covariate.
- $\Pr(Z = 1 \,|\, \mathbf{x}) = e(\mathbf{x})$, say, is the **propensity score**, and (1) $\Longrightarrow$ ignorable given $e(\mathbf{x})$.
- $\Pr(Z = 1 \,|\, \mathbf{x},\, r_T,\, r_C) = \zeta$ is the **principal unobserved covariate**.

Ignorable Treatment Assignment & Principal Unobserved Covariate

- Treatment assignment is **ignorable given the observed covariates $\mathbf{x}$** if

$$0 < \Pr(Z = 1 \,|\, \mathbf{x},\, r_T,\, r_C) = \Pr(Z = 1 \,|\, \mathbf{x}) < 1 \qquad (1)$$

- If ignorable, **adjustments** for $\mathbf{x}$ **suffice** for causal inference.
- We often speak of **ignorable assignment given something else**, given a function $\mathbf{h}(\mathbf{x})$ of $\mathbf{x}$, or given $(\mathbf{x},\, u)$ where u is an unobserved covariate.
- $\Pr(Z = 1 \,|\, \mathbf{x}) = e(\mathbf{x})$, say, is the **propensity score**, and (1) $\Longrightarrow$ ignorable given $e(\mathbf{x})$.
- $\Pr(Z = 1 \,|\, \mathbf{x},\, r_T,\, r_C) = \zeta$ is the **principal unobserved covariate**.
- Suppose $0 < \zeta < 1$. **Two key facts follow**. Then, **(i)** treatment assignment is ignorable given $\mathbf{x} \iff e(\mathbf{x}) = \zeta$, and **(ii)** treatment assignment is always ignorable given $\{\mathbf{h}(\mathbf{x}),\, \zeta\}$ for any function $\mathbf{h}(\cdot)$.

Ignorable Treatment Assignment & Principal Unobserved Covariate

- Treatment assignment is **ignorable given the observed covariates $\mathbf{x}$** if

$$0 < \Pr(Z = 1 \,|\, \mathbf{x},\, r_T,\, r_C) = \Pr(Z = 1 \,|\, \mathbf{x}) < 1 \qquad (1)$$

- If ignorable, **adjustments** for $\mathbf{x}$ **suffice** for causal inference.
- We often speak of **ignorable assignment given something else**, given a function $\mathbf{h}(\mathbf{x})$ of $\mathbf{x}$, or given $(\mathbf{x},\, u)$ where u is an unobserved covariate.
- $\Pr(Z = 1 \,|\, \mathbf{x}) = e(\mathbf{x})$, say, is the **propensity score**, and (1) $\implies$ ignorable given $e(\mathbf{x})$.
- $\Pr(Z = 1 \,|\, \mathbf{x},\, r_T,\, r_C) = \zeta$ is the **principal unobserved covariate**.
- Suppose $0 < \zeta < 1$. **Two key facts follow**. Then, **(i)** treatment assignment is ignorable given $\mathbf{x} \iff e(\mathbf{x}) = \zeta$, and **(ii)** treatment assignment is always ignorable given $\{\mathbf{h}(\mathbf{x}),\, \zeta\}$ for any function $\mathbf{h}(\cdot)$.
- Importantly, $\zeta = \Pr(Z = 1 \,|\, r_T,\, r_C,\, \mathbf{x})$ is **a function of** $(r_T,\, r_C,\, \mathbf{x})$.

Observational Block Design

- Build I blocks, $i = 1, \ldots, I$, and J people per block, $j = 1, \ldots, J$, with one treated individual per block, $1 = \sum_{j=1}^{J} Z_{ij}$ for each i.

Observational Block Design

- Build I blocks, $i = 1, \ldots, I$, and J people per block, $j = 1, \ldots, J$, with one treated individual per block, $1 = \sum_{j=1}^{J} Z_{ij}$ for each i.
- Sample independent $(R, Z, \mathbf{x})$ and assemble into blocks.

Observational Block Design

- Build I blocks, $i = 1, \ldots, I$, and J people per block, $j = 1, \ldots, J$, with one treated individual per block, $1 = \sum_{j=1}^{J} Z_{ij}$ for each i.
- Sample independent $(R,\ Z, \mathbf{x})$ and assemble into blocks.
- Create I non-overlapping blocks matched for $\mathbf{h}(\mathbf{x})$,

$$\mathbf{h}(\mathbf{x}_{i1}) = \cdots = \mathbf{h}(\mathbf{x}_{iJ}),\ \ i = 1, \ldots, I.$$

Observational Block Design

- Build I blocks, $i = 1, \ldots, I$, and J people per block, $j = 1, \ldots, J$, with one treated individual per block, $1 = \sum_{j=1}^{J} Z_{ij}$ for each i.
- Sample independent $(R,\ Z, \mathbf{x})$ and assemble into blocks.
- Create I non-overlapping blocks matched for $\mathbf{h}(\mathbf{x})$,

$$\mathbf{h}(\mathbf{x}_{i1}) = \cdots = \mathbf{h}(\mathbf{x}_{iJ}),\ \ i = 1, \ldots, I.$$

- Our **worry** is that the **blocking has not controlled the principal unobserved covariate**, ζ, so that $\zeta_{ij} \neq \zeta_{ij'}$ for some i, j.

Observational Block Design

- Build I blocks, $i = 1, \ldots, I$, and J people per block, $j = 1, \ldots, J$, with one treated individual per block, $1 = \sum_{j=1}^{J} Z_{ij}$ for each i.
- Sample independent $(R,\ Z, \mathbf{x})$ and assemble into blocks.
- Create I non-overlapping blocks matched for $\mathbf{h}(\mathbf{x})$,

$$\mathbf{h}(\mathbf{x}_{i1}) = \cdots = \mathbf{h}(\mathbf{x}_{iJ}),\ \ i = 1, \ldots, I.$$

- Our **worry** is that the **blocking has not controlled the principal unobserved covariate**, ζ, so that $\zeta_{ij} \neq \zeta_{ij'}$ for some i, j.
- **Could happen in any of three ways**: (i) controlling for $\mathbf{h}(\mathbf{x})$ did not control for $e(\mathbf{x})$, (ii) controlling for $\mathbf{h}(\mathbf{x})$ did not control for ζ_{ij} because treatment assignment is not ignorable given $\mathbf{x}$, or (iii) both (i) and (ii).

Notation

- Write $\mathcal{F} = \{(r_{Tij}, r_{Cij}, \mathbf{x}_{ij}), i = 1, \ldots, I, j = 1, \ldots, J\}$.

Notation

- Write $\mathcal{F} = \{\left(r_{Tij},\, r_{Cij},\, \mathbf{x}_{ij}\right),\, i = 1, \ldots,\, I,\, j = 1, \ldots,\, J\}$.
- Principal unobserved covarate $\zeta = \Pr(Z = 1 | r_T, r_C, \mathbf{X})$ is a function of $(r_T, r_C, \mathbf{X})$, all of which are in $\mathcal{F}$, so $\zeta_{ij} = \Pr(Z_{ij} = 1 \mid \mathcal{F})$.

Notation

- Write $\mathcal{F} = \{\left(r_{Tij},\, r_{Cij},\, \mathbf{x}_{ij}\right),\, i = 1, \ldots,\, I,\, j = 1, \ldots,\, J\}$.
- Principal unobserved covarate $\zeta = \Pr(Z = 1 | r_T, r_C, \mathbf{X})$ is a function of $(r_T, r_C, \mathbf{X})$, all of which are in $\mathcal{F}$, so $\zeta_{ij} = \Pr(Z_{ij} = 1 \,|\, \mathcal{F})$.
- Let $\mathcal{Z}$ be the set of possible values, $\mathbf{z}$, of $\mathbf{Z} = (Z_{11}, \ldots, Z_{IJ})$, so $z_{ij} = 0$ or 1, and $1 = \sum_{j=1}^{J} z_{ij}$ for $i = 1, \ldots, I$. So, $\mathcal{Z}$ contains J^I elements $\mathbf{z}$.

Notation

- Write $\mathcal{F} = \{\left(r_{Tij},\, r_{Cij},\, \mathbf{x}_{ij}\right),\, i = 1, \ldots,\, I,\, j = 1, \ldots,\, J\}$.
- Principal unobserved covarate $\zeta = \Pr(Z = 1 | r_T, r_C, \mathbf{X})$ is a function of $(r_T, r_C, \mathbf{X})$, all of which are in $\mathcal{F}$, so $\zeta_{ij} = \Pr(Z_{ij} = 1 \,|\, \mathcal{F})$.
- Let $\mathcal{Z}$ be the set of possible values, $\mathbf{z}$, of $\mathbf{Z} = (Z_{11}, \ldots, Z_{IJ})$, so $z_{ij} = 0$ or 1, and $1 = \sum_{j=1}^{J} z_{ij}$ for $i = 1, \ldots, I$. So, $\mathcal{Z}$ contains J^I elements $\mathbf{z}$.
- We sampled independent people and blocked so that $\mathbf{Z} \in \mathcal{Z}$, i.e., by conditioning on the event $\mathbf{Z} \in \mathcal{Z}$.

Notation

- Write $\mathcal{F} = \{\left(r_{Tij},\, r_{Cij},\, \mathbf{x}_{ij}\right),\, i = 1, \ldots,\, I,\, j = 1, \ldots,\, J\}$.
- Principal unobserved covarate $\zeta = \Pr(Z = 1 | r_T, r_C, \mathbf{X})$ is a function of $(r_T, r_C, \mathbf{X})$, all of which are in $\mathcal{F}$, so $\zeta_{ij} = \Pr(Z_{ij} = 1 \,|\, \mathcal{F})$.
- Let $\mathcal{Z}$ be the set of possible values, $\mathbf{z}$, of $\mathbf{Z} = (Z_{11}, \ldots, Z_{IJ})$, so $z_{ij} = 0$ or 1, and $1 = \sum_{j=1}^{J} z_{ij}$ for $i = 1, \ldots, I$. So, $\mathcal{Z}$ contains J^I elements $\mathbf{z}$.
- We sampled independent people and blocked so that $\mathbf{Z} \in \mathcal{Z}$, i.e., by conditioning on the event $\mathbf{Z} \in \mathcal{Z}$.
- Abbreviate conditioning on $\mathbf{Z} \in \mathcal{Z}$ as conditioning on $\mathcal{Z}$.

Notation

- Write $\mathcal{F} = \{\left(r_{Tij},\, r_{Cij},\, \mathbf{x}_{ij}\right),\, i = 1, \ldots,\, I,\, j = 1, \ldots,\, J\}$.
- Principal unobserved covarate $\zeta = \Pr(Z = 1 | r_T, r_C, \mathbf{X})$ is a function of $(r_T, r_C, \mathbf{X})$, all of which are in $\mathcal{F}$, so $\zeta_{ij} = \Pr(Z_{ij} = 1 \,|\, \mathcal{F})$.
- Let $\mathcal{Z}$ be the set of possible values, $\mathbf{z}$, of $\mathbf{Z} = (Z_{11}, \ldots, Z_{IJ})$, so $z_{ij} = 0$ or 1, and $1 = \sum_{j=1}^{J} z_{ij}$ for $i = 1, \ldots, I$. So, $\mathcal{Z}$ contains J^I elements $\mathbf{z}$.
- We sampled independent people and blocked so that $\mathbf{Z} \in \mathcal{Z}$, i.e., by conditioning on the event $\mathbf{Z} \in \mathcal{Z}$.
- Abbreviate conditioning on $\mathbf{Z} \in \mathcal{Z}$ as conditioning on $\mathcal{Z}$.
- For example, in a randomized block design,

$$\frac{1}{J} = \Pr(Z_{ij} = 1 \,|\, \mathcal{F}, \mathcal{Z})$$

Bias Within Blocks; Introducing θ_{ij}

- Given $\mathcal{F}$, the chance that ij is the only treated individual in block i is the chance that $Z_{ij} = 1$ and $Z_{ik} = 0$ for $k \neq i$

$$\zeta_{ij} \prod_{k \neq j}^{J} (1 - \zeta_{ik}) = \frac{\zeta_{ij}}{1 - \zeta_{ij}} \prod_{k=1}^{J} (1 - \zeta_{ik}),$$

Bias Within Blocks; Introducing θ_{ij}

- Given $\mathcal{F}$, the chance that ij is the only treated individual in block i is the chance that $Z_{ij} = 1$ and $Z_{ik} = 0$ for $k \neq i$

$$\zeta_{ij} \prod_{k \neq j}^{J} (1 - \zeta_{ik}) = \frac{\zeta_{ij}}{1 - \zeta_{ij}} \prod_{k=1}^{J} (1 - \zeta_{ik}),$$

- So, conditioning on $\sum_{k=1}^{J} Z_{ik} = 1$ says $\Pr\left(Z_{ij} = 1 \,|r_{Tij}, r_{Cij}, \mathbf{x}_{ij}, \sum Z_{ik} = 1\right)$ equals

$$\Pr\left(Z_{ij} = 1 \,|\mathcal{F}, \mathcal{Z}\right) = \frac{\frac{\zeta_{ij}}{1-\zeta_{ij}}}{\sum_{k=1}^{J} \frac{\zeta_{ik}}{1-\zeta_{ik}}} = \theta_{ij},$$

  say, where $1 = \sum_{j=1}^{J} \theta_{ij}$ for each i.

Sensitivity Analysis in Terms of ζ

- From the previous slide, $1 = \sum_{j=1}^{J} \theta_{ij}$ and

$$\Pr\left(Z_{ij} = 1 \,\middle|\, r_{Tij}, r_{Cij}, \mathbf{x}_{ij}, \sum Z_{ik} = 1\right) = \Pr\left(Z_{ij} = 1 \,|\, \mathcal{F}, \mathcal{Z}\right) = \frac{\frac{\zeta_{ij}}{1-\zeta_{ij}}}{\sum_{k=1}^{J} \frac{\zeta_{ik}}{1-\zeta_{ik}}} = \theta_{ij},$$

Sensitivity Analysis in Terms of ζ

- From the previous slide, $1 = \sum_{j=1}^{J} \theta_{ij}$ and

$$\Pr\left(Z_{ij} = 1 \,\middle|\, r_{Tij}, r_{Cij}, \mathbf{x}_{ij}, \sum Z_{ik} = 1\right) = \Pr\left(Z_{ij} = 1 \,|\, \mathcal{F}, \mathcal{Z}\right) = \frac{\frac{\zeta_{ij}}{1-\zeta_{ij}}}{\sum_{k=1}^{J} \frac{\zeta_{ik}}{1-\zeta_{ik}}} = \theta_{ij},$$

- Sensitivity analysis in terms of the principal unobserved covariate $\zeta = \Pr(Z = 1 \,|\, r_t, r_C, \mathbf{x})$

$$\Gamma \geq \frac{\zeta_{ij}\left(1-\zeta_{ij'}\right)}{\zeta_{ij'}\left(1-\zeta_{ij}\right)} \geq \frac{1}{\Gamma} \text{ for all } i, j, j'.$$

is the same as

$$\Gamma \geq \frac{\theta_{ij}}{\theta_{ij'}} \geq \frac{1}{\Gamma} \text{ for all } i, j, j'$$

Comparing Methods and Designs for Observational Studies

- **Different statistics, different research designs, correctly yield different levels of sensitivity to unobserved biases**.

Comparing Methods and Designs for Observational Studies

- **Different statistics, different research designs, correctly yield different levels of sensitivity to unobserved biases**.
- We would like to understand this, so we can make wise choices in design and analysis.

Comparing Methods and Designs for Observational Studies

- **Different statistics, different research designs, correctly yield different levels of sensitivity to unobserved biases**.
- We would like to understand this, so we can make wise choices in design and analysis.
- **First**, let's do an **analysis of the alcohol data** and see it happen in one data set.

Comparing Methods and Designs for Observational Studies

- **Different statistics, different research designs, correctly yield different levels of sensitivity to unobserved biases**.
- We would like to understand this, so we can make wise choices in design and analysis.
- **First**, let's do an **analysis of the alcohol data** and see it happen in one data set.
- **Second**, set aside our one data set, **replace it by a probability model that generates data**, and demonstrate that what happened once in data should always happen, measuring precisely when and to what degree it happens.

Comparing Methods and Designs for Observational Studies

- **Different statistics, different research designs, correctly yield different levels of sensitivity to unobserved biases**.
- We would like to understand this, so we can make wise choices in design and analysis.
- **First**, let's do an **analysis of the alcohol data** and see it happen in one data set.
- **Second**, set aside our one data set, **replace it by a probability model that generates data**, and demonstrate that what happened once in data should always happen, measuring precisely when and to what degree it happens.
- Start with a **collection of closely related statistics**, including familiar and unfamiliar statistics. See how the results vary in this collection.

Weighted Rank Statistics

- Test the hypothesis of no effect, $H_0 : r_{Tij} = r_{Cij}, \forall i, j.$

Weighted Rank Statistics

- Test the hypothesis of no effect, $H_0 : r_{Tij} = r_{Cij}, \forall i, j$.
- Rank R_{ij} from 1 to J in each block i, with average ranks for ties, calling the within-block ranks q_{ij}.

Weighted Rank Statistics

- Test the hypothesis of no effect, $H_0 : r_{Tij} = r_{Cij}, \forall i, j$.
- Rank R_{ij} from 1 to J in each block i, with average ranks for ties, calling the within-block ranks q_{ij}.
- Let Iw_i be the rank the ith of the I within-block ranges $b_i = \max R_{ij} - \min R_{ij}$, with average ranks for ties, so $0 \leq w_i \leq 1$.

Weighted Rank Statistics

- Test the hypothesis of no effect, $H_0 : r_{Tij} = r_{Cij}, \forall i, j$.
- Rank R_{ij} from 1 to J in each block i, with average ranks for ties, calling the within-block ranks q_{ij}.
- Let Iw_i be the rank the ith of the I within-block ranges $b_i = \max R_{ij} - \min R_{ij}$, with average ranks for ties, so $0 \leq w_i \leq 1$.
- Score the ranks of the ranges by a function $\varphi(w_i)$, where $\varphi : [0, 1] \rightarrow [0, 1]$.

Weighted Rank Statistics

- Test the hypothesis of no effect, $H_0 : r_{Tij} = r_{Cij}, \forall i, j$.
- Rank R_{ij} from 1 to J in each block i, with average ranks for ties, calling the within-block ranks q_{ij}.
- Let Iw_i be the rank the ith of the I within-block ranges $b_i = \max R_{ij} - \min R_{ij}$, with average ranks for ties, so $0 \leq w_i \leq 1$.
- Score the ranks of the ranges by a function $\varphi(w_i)$, where $\varphi : [0, 1] \rightarrow [0, 1]$.
- The test statistic is $T = \sum_i^I \varphi(w_i) \sum_{j=1}^J Z_{ij}\, q_{ij}$.

Weighted Rank Statistics

- Test the hypothesis of no effect, $H_0 : r_{Tij} = r_{Cij}, \forall i, j$.
- Rank R_{ij} from 1 to J in each block i, with average ranks for ties, calling the within-block ranks q_{ij}.
- Let Iw_i be the rank the ith of the I within-block ranges $b_i = \max R_{ij} - \min R_{ij}$, with average ranks for ties, so $0 \leq w_i \leq 1$.
- Score the ranks of the ranges by a function $\varphi(w_i)$, where $\varphi : [0, 1] \to [0, 1]$.
- The test statistic is $T = \sum_i^I \varphi(w_i) \sum_{j=1}^J Z_{ij}\, q_{ij}$.
- For pairs, $J = 2$, taking $\varphi(w) = 1$ yields the sign test, taking $\varphi(w) = w$ yields Wilcoxon's signed rank test, and for general $\varphi(w)$ it is a general signed rank test.

Weighted Rank Statistics

- Test the hypothesis of no effect, $H_0 : r_{Tij} = r_{Cij}, \forall i, j$.
- Rank R_{ij} from 1 to J in each block i, with average ranks for ties, calling the within-block ranks q_{ij}.
- Let Iw_i be the rank the ith of the I within-block ranges $b_i = \max R_{ij} - \min R_{ij}$, with average ranks for ties, so $0 \leq w_i \leq 1$.
- Score the ranks of the ranges by a function $\varphi(w_i)$, where $\varphi : [0, 1] \rightarrow [0, 1]$.
- The test statistic is $T = \sum_i^I \varphi(w_i) \sum_{j=1}^J Z_{ij}\, q_{ij}$.
- For pairs, $J = 2$, taking $\varphi(w) = 1$ yields the sign test, taking $\varphi(w) = w$ yields Wilcoxon's signed rank test, and for general $\varphi(w)$ it is a general signed rank test.
- For $J \geq 2$, taking $\varphi(w_i) = 1$ yields the blocked Wilcoxon rank sum test (Lehmann 1975 *Nonparametrics*, §3.3), and taking $\varphi(w) = w$ yields Quade's (1979, JASA) statistic.

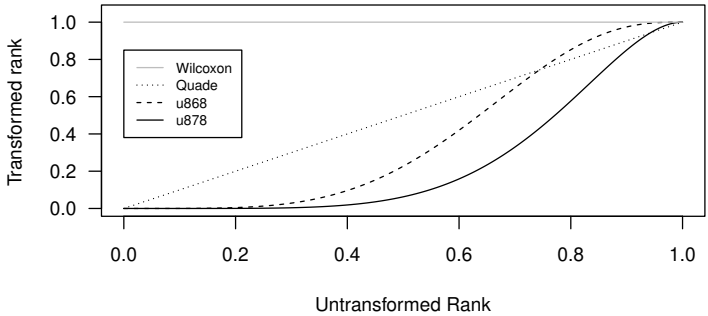

Figure: Four weight functions $\varphi(w)$ of the block ranges.

The Set B_Γ of Biased Treatment Assignments $\boldsymbol{\theta}$

- Define B_Γ as the set of all $\boldsymbol{\theta} = (\theta_{11}, \ldots, \theta_{IJ})$ such that:

$$1 = \sum_j \theta_{ij},\ i = 1, \ldots, I \quad \text{and} \quad \Gamma \geq \frac{\theta_{ij}}{\theta_{ij'}} \geq \frac{1}{\Gamma} \text{ for all } i, j, j'$$

The Set B_Γ of Biased Treatment Assignments $\boldsymbol{\theta}$

- Define B_Γ as the set of all $\boldsymbol{\theta} = (\theta_{11}, \ldots, \theta_{IJ})$ such that:

$$1 = \sum_j \theta_{ij},\ i = 1, \ldots, I \quad \text{and} \quad \Gamma \geq \frac{\theta_{ij}}{\theta_{ij'}} \geq \frac{1}{\Gamma} \text{ for all } i, j, j'$$

- With $I = 406$ and $J = 4$, each $\boldsymbol{\theta}$ is of **dimension** $IJ = 1624$ but lives in **flat of dimension** $I(J-1) = 1218$. B_Γ is a closed and bounded (hence compact) set of $\boldsymbol{\theta}$'s.

The Set B_Γ of Biased Treatment Assignments $\boldsymbol{\theta}$

- Define B_Γ as the set of all $\boldsymbol{\theta} = (\theta_{11}, \ldots, \theta_{IJ})$ such that:

$$1 = \sum_j \theta_{ij},\ i = 1, \ldots, I \quad \text{and} \quad \Gamma \geq \frac{\theta_{ij}}{\theta_{ij'}} \geq \frac{1}{\Gamma} \text{ for all } i, j, j'$$

- With $I = 406$ and $J = 4$, each $\boldsymbol{\theta}$ is of **dimension** $IJ = 1624$ but lives in **flat of dimension** $I(J-1) = 1218$. B_Γ is a closed and bounded (hence compact) set of $\boldsymbol{\theta}$'s.
- **Nested sets**, $B_\Gamma \subset B_{\Gamma'}$ for $\Gamma < \Gamma'$, assume less and less as $\Gamma \to \infty$.

The Set B_Γ of Biased Treatment Assignments $\boldsymbol{\theta}$

- Define B_Γ as the set of all $\boldsymbol{\theta} = (\theta_{11}, \ldots, \theta_{IJ})$ such that:

$$1 = \sum_j \theta_{ij},\ i = 1, \ldots, I \quad \text{and} \quad \Gamma \geq \frac{\theta_{ij}}{\theta_{ij'}} \geq \frac{1}{\Gamma} \text{ for all } i, j, j'$$

- With $I = 406$ and $J = 4$, each $\boldsymbol{\theta}$ is of **dimension** $IJ = 1624$ but lives in **flat of dimension** $I(J-1) = 1218$. B_Γ is a closed and bounded (hence compact) set of $\boldsymbol{\theta}$'s.
- **Nested sets**, $B_\Gamma \subset B_{\Gamma'}$ for $\Gamma < \Gamma'$, assume less and less as $\Gamma \to \infty$.
- Every $\boldsymbol{\theta}$ with $0 < \theta_{ij} < 1$ and $1 = \sum_j \theta_{ij}$ is in some B_Γ for large enough Γ.

The Set B_Γ of Biased Treatment Assignments $\boldsymbol{\theta}$

- Define B_Γ as the set of all $\boldsymbol{\theta} = (\theta_{11}, \ldots, \theta_{IJ})$ such that:

$$1 = \sum_j \theta_{ij},\ i = 1, \ldots, I \quad \text{and} \quad \Gamma \geq \frac{\theta_{ij}}{\theta_{ij'}} \geq \frac{1}{\Gamma} \text{ for all } i, j, j'$$

- With $I = 406$ and $J = 4$, each $\boldsymbol{\theta}$ is of **dimension** $IJ = 1624$ but lives in **flat of dimension** $I(J-1) = 1218$. B_Γ is a closed and bounded (hence compact) set of $\boldsymbol{\theta}$'s.
- **Nested sets**, $B_\Gamma \subset B_{\Gamma'}$ for $\Gamma < \Gamma'$, assume less and less as $\Gamma \to \infty$.
- Every $\boldsymbol{\theta}$ with $0 < \theta_{ij} < 1$ and $1 = \sum_j \theta_{ij}$ is in some B_Γ for large enough Γ.
- A **randomized block design** has $\boldsymbol{\theta} = \overline{\boldsymbol{\theta}}$ where $\overline{\theta}_{ij} = 1/J$ or equivalently $\boldsymbol{\theta} \in B_1$.

The Set B_Γ of Biased Treatment Assignments $\boldsymbol{\theta}$

- Define B_Γ as the set of all $\boldsymbol{\theta} = (\theta_{11}, \ldots, \theta_{IJ})$ such that:

$$1 = \sum_j \theta_{ij},\ i = 1, \ldots, I \quad \text{and} \quad \Gamma \geq \frac{\theta_{ij}}{\theta_{ij'}} \geq \frac{1}{\Gamma} \text{ for all } i, j, j'$$

- With $I = 406$ and $J = 4$, each $\boldsymbol{\theta}$ is of **dimension** $IJ = 1624$ but lives in **flat of dimension** $I(J-1) = 1218$. B_Γ is a closed and bounded (hence compact) set of $\boldsymbol{\theta}$'s.
- **Nested sets**, $B_\Gamma \subset B_{\Gamma'}$ for $\Gamma < \Gamma'$, assume less and less as $\Gamma \to \infty$.
- Every $\boldsymbol{\theta}$ with $0 < \theta_{ij} < 1$ and $1 = \sum_j \theta_{ij}$ is in some B_Γ for large enough Γ.
- A **randomized block design** has $\boldsymbol{\theta} = \overline{\boldsymbol{\theta}}$ where $\overline{\theta}_{ij} = 1/J$ or equivalently $\boldsymbol{\theta} \in B_1$.
- The **central problem in an observational block design** is that there is no basis for assuming $\boldsymbol{\theta} \in B_1$. For $\Gamma > 1$, $\boldsymbol{\theta} \in B_\Gamma$ does not identify $\boldsymbol{\theta}$.

Sensitivity Analysis

- Reject H_0 if $T \geq t$ where $T = \sum_i \varphi(w_i) \sum_j Z_{ij}\, q_{ij}$ and $\Pr(\mathbf{Z} = \mathbf{z} \,|\mathcal{F}, \mathcal{Z}) = \prod_i \prod_j \theta_{ij}^{z_{ij}}$.

Sensitivity Analysis

- Reject H_0 if $T \geq t$ where $T = \sum_i \varphi(w_i) \sum_j Z_{ij}\, q_{ij}$ and $\Pr(\mathbf{Z} = \mathbf{z} \,|\mathcal{F}, \mathcal{Z}) = \prod_i \prod_j \theta_{ij}^{z_{ij}}$.
- $[A] = 1$ if event A occurs; otherwise 0. Rejection of H_0: $\left[\sum_i \varphi(w_i) \sum_j z_{ij}\, q_{ij} \geq t\right] = 1$.

Sensitivity Analysis

- Reject H_0 if $T \geq t$ where $T = \sum_i \varphi(w_i) \sum_j Z_{ij}\, q_{ij}$ and $\Pr(\mathbf{Z} = \mathbf{z} \,|\mathcal{F}, \mathcal{Z}) = \prod_i \prod_j \theta_{ij}^{z_{ij}}$.
- $[A] = 1$ if event A occurs; otherwise 0. Rejection of H_0: $\left[\sum_i \varphi(w_i) \sum_j z_{ij}\, q_{ij} \geq t\right] = 1$.
- **For fixed θ, rejection occurs with probability**

$$\sum_{\mathbf{z} \in \mathcal{Z}} \left[\sum_i \varphi(w_i) \sum_j z_{ij}\, q_{ij} \geq t\right] \prod_i \prod_j \theta_{ij}^{z_{ij}}$$

Sensitivity Analysis

- Reject H_0 if $T \geq t$ where $T = \sum_i \varphi(w_i) \sum_j Z_{ij}\, q_{ij}$ and $\Pr(\mathbf{Z} = \mathbf{z} \,|\mathcal{F}, \mathcal{Z}) = \prod_i \prod_j \theta_{ij}^{z_{ij}}$.
- $[A] = 1$ if event A occurs; otherwise 0. Rejection of H_0: $\left[\sum_i \varphi(w_i) \sum_j z_{ij}\, q_{ij} \geq t\right] = 1$.
- **For fixed θ, rejection occurs with probability**

$$\sum_{\mathbf{z} \in \mathcal{Z}} \left[\sum_i \varphi(w_i) \sum_j z_{ij}\, q_{ij} \geq t\right] \prod_i \prod_j \theta_{ij}^{z_{ij}}$$

- For a given $\Gamma \geq 1$, **the max P-value** for $\boldsymbol{\theta} \in B_\Gamma$ is

$$P_\Gamma = \max_{\boldsymbol{\theta} \in B_\Gamma} \sum_{\mathbf{z} \in \mathcal{Z}} \left[\sum_i \varphi(w_i) \sum_j z_{ij}\, q_{ij} \geq t\right] \prod_i \prod_j \theta_{ij}^{z_{ij}}$$

Sensitivity Analysis, Alcohol Example, Comparing 4 Statistics

Table: Upper bounds on one-sided P-values testing no effect of light daily alcohol on HDL Cholesterol. In a column, **bold** is a P-value near 0.05. Hammond's (1964, *JNCI*) study of smoking and lung cancer is sensitive to a bias of $\Gamma = 6$. The choice of test statistic matters.

Γ	Wilcoxon	Quade	U868	U878
1	0.0000	0.0000	0.0000	0.0000
2	0.0000	0.0000	0.0000	0.0000
3.5	**0.0603**	0.0002	0.0000	0.0000
4	0.3478	0.0052	0.0003	0.0001
4.5	0.7401	**0.0447**	0.0028	0.0010
5	0.9429	0.1775	0.0154	0.0050
5.5	0.9926	0.4123	**0.0537**	0.0174
6	0.9994	0.6642	0.1340	**0.0456**

Is 1-to-3 Better Than 1-to-1? A Fair Comparison

- Want to see in the data my earlier claim that 1-to-3 blocks more insensitive to bias than 1-to-1 pairs.

Is 1-to-3 Better Than 1-to-1? A Fair Comparison

- Want to see in the data my earlier claim that 1-to-3 blocks more insensitive to bias than 1-to-1 pairs.
- Not fair to compare 406 1-to-3 blocks to 406 1-to-1 pairs.

Is 1-to-3 Better Than 1-to-1? A Fair Comparison

- Want to see in the data my earlier claim that 1-to-3 blocks more insensitive to bias than 1-to-1 pairs.
- Not fair to compare 406 1-to-3 blocks to 406 1-to-1 pairs.
- Consider the usual Gaussian linear model, additive block effects, constant within block variance σ^2. Estimator is the mean of the treated-minus-average control difference.

Is 1-to-3 Better Than 1-to-1? A Fair Comparison

- Want to see in the data my earlier claim that 1-to-3 blocks more insensitive to bias than 1-to-1 pairs.
- Not fair to compare 406 1-to-3 blocks to 406 1-to-1 pairs.
- Consider the usual Gaussian linear model, additive block effects, constant within block variance σ^2. Estimator is the mean of the treated-minus-average control difference.
- With M 1-to-1 pairs, estimator has variance $2\sigma^2/M$. With I 1-to-3 blocks, estimator has variance $(1 + 1/3)\sigma^2/I$. As far as the standard error goes, M pairs is about the same as I 1-to-3 blocks if $I = (1 + 1/3)M/2$. For $M = 406$ pairs, take $I = 2M/3 \doteq 271$ blocks.

Fair Comparison, Pairs Versus Blocks

Table: Bounds on P-values for the hypothesis of no effect. Last P-value ≤ 0.05 is in **bold**.

	406 1-to-1 Pairs				271 1-to-3 Blocks			
Γ	Wilcoxon	Quade	U868	U878	Wilcoxon	Quade	U868	U878
1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	**0.006**	**0.000**	0.000	0.000	0.000	0.000	0.000	0.000
3.5	0.994	0.233	**0.013**	0.003	**0.044**	0.001	0.000	0.000
4	1.000	0.584	0.064	0.015	0.224	0.008	0.001	0.001
4.5	1.000	0.851	0.182	**0.046**	0.532	**0.045**	0.007	0.004
5	1.000	0.963	0.359	0.106	0.799	0.143	**0.024**	0.014
5.5	1.000	0.993	0.552	0.198	0.937	0.310	0.063	**0.034**
6	1.000	0.999	0.720	0.311	0.985	0.511	0.131	0.069

Can We Understand This Theoretically?

- Suppose that we have some **model that generated the data**, I blocks of size J, one treated individual per block. Let $I \to \infty$.

Can We Understand This Theoretically?

- Suppose that we have some **model that generated the data**, I blocks of size J, one treated individual per block. Let $I \to \infty$.
- **When letting** $I \to \infty$, quantities gain a **subscript** I: T becomes T_I, for example.

Can We Understand This Theoretically?

- Suppose that we have some **model that generated the data**, I blocks of size J, one treated individual per block. Let $I \to \infty$.
- **When letting** $I \to \infty$, quantities gain a **subscript** I: T becomes T_I, for example.
- **A basic block model** with continuous & bivariate exchangeable errors $(\varepsilon_{Tij}, \varepsilon_{Cij})$

$$r_{Tij} = \mu + \beta_i + \tau + \varepsilon_{Tij}, \qquad r_{Cij} = \mu + \beta_i + \varepsilon_{Cij}, \qquad r_{Tij} - r_{Cij} = \tau + \varepsilon_{Tij} - \varepsilon_{Cij},$$

Can We Understand This Theoretically?

- Suppose that we have some **model that generated the data**, I blocks of size J, one treated individual per block. Let $I \to \infty$.
- **When letting** $I \to \infty$, quantities gain a **subscript** I: T becomes T_I, for example.
- **A basic block model** with continuous & bivariate exchangeable errors $(\varepsilon_{Tij}, \varepsilon_{Cij})$

$$r_{Tij} = \mu + \beta_i + \tau + \varepsilon_{Tij}, \qquad r_{Cij} = \mu + \beta_i + \varepsilon_{Cij}, \qquad r_{Tij} - r_{Cij} = \tau + \varepsilon_{Tij} - \varepsilon_{Cij},$$

- **Treated-minus-control pair difference** in block i is $r_{Tij} - r_{Cij'} = \tau + \varepsilon_{Tij} - \varepsilon_{Cij'}$ is symmetric about τ. Numerical work takes $1 = \sqrt{\text{var}(\varepsilon_{Tij} - \varepsilon_{Cij'})}$ and $\tau = 1/2$.

Can We Understand This Theoretically?

- Suppose that we have some **model that generated the data**, I blocks of size J, one treated individual per block. Let $I \to \infty$.
- **When letting** $I \to \infty$, quantities gain a **subscript** I: T becomes T_I, for example.
- **A basic block model** with continuous & bivariate exchangeable errors $(\varepsilon_{Tij}, \varepsilon_{Cij})$

$$r_{Tij} = \mu + \beta_i + \tau + \varepsilon_{Tij}, \qquad r_{Cij} = \mu + \beta_i + \varepsilon_{Cij}, \qquad r_{Tij} - r_{Cij} = \tau + \varepsilon_{Tij} - \varepsilon_{Cij},$$

- **Treated-minus-control pair difference** in block i is $r_{Tij} - r_{Cij'} = \tau + \varepsilon_{Tij} - \varepsilon_{Cij'}$ is symmetric about τ. Numerical work takes $1 = \sqrt{\text{var}(\varepsilon_{Tij} - \varepsilon_{Cij'})}$ and $\tau = 1/2$.
- Imagine the **study is unaffected by unmeasured bias** (i.e., ignorable given $\mathbf{x}$), so that $1/J = \theta_{ij} = \Pr\left(Z_{ij} = 1 \,\middle|\, r_{Tij}, r_{Cij}, \mathbf{x}_{ij}, \sum Z_{ik} = 1\right), \forall i, j.$

Can We Understand This Theoretically?

- Suppose that we have some **model that generated the data**, I blocks of size J, one treated individual per block. Let $I \to \infty$.
- **When letting** $I \to \infty$, quantities gain a **subscript** I: T becomes T_I, for example.
- **A basic block model** with continuous & bivariate exchangeable errors $(\varepsilon_{Tij}, \varepsilon_{Cij})$

$$r_{Tij} = \mu + \beta_i + \tau + \varepsilon_{Tij}, \qquad r_{Cij} = \mu + \beta_i + \varepsilon_{Cij}, \qquad r_{Tij} - r_{Cij} = \tau + \varepsilon_{Tij} - \varepsilon_{Cij},$$

- **Treated-minus-control pair difference** in block i is $r_{Tij} - r_{Cij'} = \tau + \varepsilon_{Tij} - \varepsilon_{Cij'}$ is symmetric about τ. Numerical work takes $1 = \sqrt{\text{var}(\varepsilon_{Tij} - \varepsilon_{Cij'})}$ and $\tau = 1/2$.
- Imagine the **study is unaffected by unmeasured bias** (i.e., ignorable given $\mathbf{x}$), so that $1/J = \theta_{ij} = \Pr\left(Z_{ij} = 1 \,\middle|\, r_{Tij}, r_{Cij}, \mathbf{x}_{ij}, \sum Z_{ik} = 1\right), \forall i, j.$
- If $\tau \neq 0$, then it is in precisely this sort of case (a so-called **favorable situation**) that you hope to report insensitivity to unmeasured biases. Will you?

Design Sensitivity and Bahadur Efficiency Under a Favorable Model

- Test no effect, H_0, with bias $\leq \Gamma$ versus a simple H_A from a favorable model.

Design Sensitivity and Bahadur Efficiency Under a Favorable Model

- Test no effect, H_0, with bias $\leq \Gamma$ versus a simple H_A from a favorable model.
- For fixed Γ, with I blocks, and a required power $\omega = 0.9$, there is a level of the test, $\alpha_{\Gamma I}$, that achieves that power. For $\Gamma = 1$, that is the level of the randomization test. For large Γ, that level might be close to 1.

Design Sensitivity and Bahadur Efficiency Under a Favorable Model

- Test no effect, H_0, with bias $\leq \Gamma$ versus a simple H_A from a favorable model.
- For fixed Γ, with I blocks, and a required power $\omega = 0.9$, there is a level of the test, $\alpha_{\Gamma I}$, that achieves that power. For $\Gamma = 1$, that is the level of the randomization test. For large Γ, that level might be close to 1.
- Want $\alpha_{\Gamma I} \to 0$ as fast as possible as $I \to \infty$. Ultimately, ω does not matter.

Design Sensitivity and Bahadur Efficiency Under a Favorable Model

- Test no effect, H_0, with bias $\leq \Gamma$ versus a simple H_A from a favorable model.
- For fixed Γ, with I blocks, and a required power $\omega = 0.9$, there is a level of the test, $\alpha_{\Gamma I}$, that achieves that power. For $\Gamma = 1$, that is the level of the randomization test. For large Γ, that level might be close to 1.
- Want $\alpha_{\Gamma I} \to 0$ as fast as possible as $I \to \infty$. Ultimately, ω does not matter.
- There is (typically) a number, $\widetilde{\Gamma}$ called the design sensitivity, such that, as $I \to \infty$:

$$\alpha_{\Gamma I} \to 0 \text{ for } \Gamma < \widetilde{\Gamma}, \qquad \alpha_{\Gamma I} \to 1 \text{ for } \Gamma > \widetilde{\Gamma}.$$

Design Sensitivity and Bahadur Efficiency Under a Favorable Model

- Test no effect, H_0, with bias $\leq \Gamma$ versus a simple H_A from a favorable model.
- For fixed Γ, with I blocks, and a required power $\omega = 0.9$, there is a level of the test, $\alpha_{\Gamma I}$, that achieves that power. For $\Gamma = 1$, that is the level of the randomization test. For large Γ, that level might be close to 1.
- Want $\alpha_{\Gamma I} \to 0$ as fast as possible as $I \to \infty$. Ultimately, ω does not matter.
- There is (typically) a number, $\widetilde{\Gamma}$ called the design sensitivity, such that, as $I \to \infty$:

$$\alpha_{\Gamma I} \to 0 \text{ for } \Gamma < \widetilde{\Gamma}, \qquad \alpha_{\Gamma I} \to 1 \text{ for } \Gamma > \widetilde{\Gamma}.$$

- If $\Gamma < \widetilde{\Gamma}$ there is (typically) a Bahadur slope $\rho_\Gamma/2 > 0$ such that

$$\rho_\Gamma = -\lim_{I\to\infty} \frac{\log(\alpha_{\Gamma I})}{I} \quad \text{so that} \quad \alpha_{\Gamma I} \approx \exp(-I\rho_\Gamma) \text{ as } I \to \infty.$$

Design Sensitivity and Bahadur Efficiency Under a Favorable Model

- Test no effect, H_0, with bias $\leq \Gamma$ versus a simple H_A from a favorable model.
- For fixed Γ, with I blocks, and a required power $\omega = 0.9$, there is a level of the test, $\alpha_{\Gamma I}$, that achieves that power. For $\Gamma = 1$, that is the level of the randomization test. For large Γ, that level might be close to 1.
- Want $\alpha_{\Gamma I} \to 0$ as fast as possible as $I \to \infty$. Ultimately, ω does not matter.
- There is (typically) a number, $\widetilde{\Gamma}$ called the design sensitivity, such that, as $I \to \infty$:

$$\alpha_{\Gamma I} \to 0 \text{ for } \Gamma < \widetilde{\Gamma}, \qquad \alpha_{\Gamma I} \to 1 \text{ for } \Gamma > \widetilde{\Gamma}.$$

- If $\Gamma < \widetilde{\Gamma}$ there is (typically) a Bahadur slope $\rho_\Gamma/2 > 0$ such that

$$\rho_\Gamma = -\lim_{I\to\infty} \frac{\log(\alpha_{\Gamma I})}{I} \quad \text{so that} \quad \alpha_{\Gamma I} \approx \exp(-I\rho_\Gamma) \text{ as } I \to \infty.$$

- The ratio of two Bahadur slopes is the Bahadur (1960) relative efficiency. Better than Pitman efficiency for observational studies because Pitman lets $\tau \to 0$ as $I \to \infty$.

Some Design Sensitivities

■ What you saw in the example happens in the limit as $I \to \infty$ for the block model with Normal errors.

Table: Design sensitivity $\widetilde{\Gamma}$ with Normal errors and $\tau = 1/2$ of the standard deviation of a treated-minus-control pair difference. The best result in each situation is in **bold**.

		Wilcoxon	Quade	U868	U878
$J = 2$	Pairs	2.2	3.2	4.2	**5.1**
$J = 4$	1-to-3 Blocks	3.5	4.4	5.2	**5.7**

■ Results for $\tau = 1/3$ have smaller $\widetilde{\Gamma}$, but a similar pattern. E.g., Quade has $\widetilde{\Gamma} = 2.1$ for $J = 2$ and $\widetilde{\Gamma} = 2.8$ for $J = 4$, while U878 has $\widetilde{\Gamma} = 2.8$ for $J = 2$ and $\widetilde{\Gamma} = 3.2$ for $J = 4$.
■ Results for $\tau = 1/2$ and errors with a t_5-distribution are similar.

Table: **Efficiency at** $\Gamma = 2$. Comparing Block Sizes $J = 2$ to $J = 4$ in a sensitivity analysis. Top half is pure block size. Bottom half is block size plus change in test statistic.

Sensitivity Analysis Performed with $\Gamma = 2$				
	$\tau = 1/2$		$\tau = 1/3$	
J	Normal	t_5	Normal	t_5
	U868 compared to U868 at $J = 2$			
2	1.00	1.00	1.00	1.00
3	1.37	1.23	2.14	1.66
4	1.83	1.63	3.07	2.29
	U868 compared to SRS at $J = 2$			
2	1.58	1.26	8.08	3.05
3	2.16	1.55	17.26	5.07
4	2.89	2.04	24.81	6.98

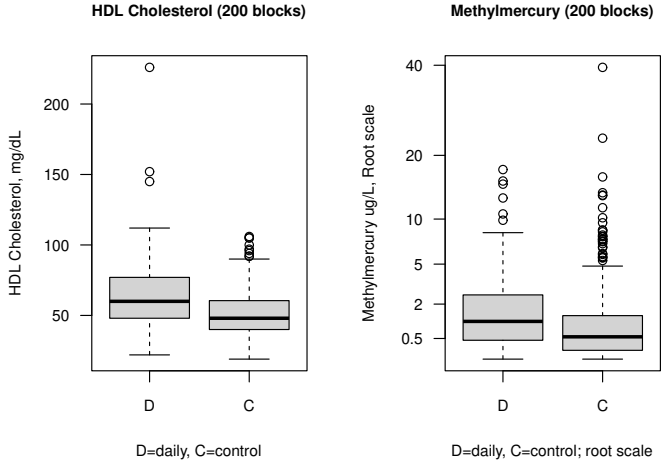

Figure: **Right panel tests the conjunction** of $H_0 : \boldsymbol{\theta} = \overline{\boldsymbol{\theta}}$ (or equivalently $H_0 : \boldsymbol{\theta} \in B_1$) and no effect on methylmercury, rejecting it with a P-value too small to calculate using Quade's statistic.

Unaffected Outcomes

- What should we make of the evidence from methylmercury of biased treatment assignment evident, $\boldsymbol{\theta} \neq \overline{\boldsymbol{\theta}}$, where $\overline{\theta}_{ij} = 1/J, \forall i,j$ and $B_1 = \{\overline{\boldsymbol{\theta}}\}$?

Unaffected Outcomes

- What should we make of the evidence from methylmercury of biased treatment assignment evident, $\boldsymbol{\theta} \neq \overline{\boldsymbol{\theta}}$, where $\overline{\theta}_{ij} = 1/J, \forall i,j$ and $B_1 = \{\overline{\boldsymbol{\theta}}\}$?
- To avoid many possible combinations, will do two-sided, 0.05-level Quade's tests for both effect and bias. (This is intended to be simple and not distracting, rather than optimal (e.g., see Table 3 in R 2023 Stat. Sci.).

Unaffected Outcomes

- What should we make of the evidence from methylmercury of biased treatment assignment evident, $\boldsymbol{\theta} \neq \overline{\boldsymbol{\theta}}$, where $\overline{\theta}_{ij} = 1/J, \forall i,j$ and $B_1 = \{\overline{\boldsymbol{\theta}}\}$?
- To avoid many possible combinations, will do two-sided, 0.05-level Quade's tests for both effect and bias. (This is intended to be simple and not distracting, rather than optimal (e.g., see Table 3 in R 2023 Stat. Sci.).
- Using the 200 blocks with methylmercury levels (rather than all 406 blocks), rejection of no effect on HDL cholesterol level becomes sensitive at $\Gamma = 3.614$. No $\boldsymbol{\theta} \in B_{3.614}$ would lead to a P-value above 0.05. (Was $B_{4.5}$ for $I = 406$.)

Unaffected Outcomes

- What should we make of the evidence from methylmercury of biased treatment assignment evident, $\boldsymbol{\theta} \neq \overline{\boldsymbol{\theta}}$, where $\overline{\theta}_{ij} = 1/J, \forall i,j$ and $B_1 = \{\overline{\boldsymbol{\theta}}\}$?
- To avoid many possible combinations, will do two-sided, 0.05-level Quade's tests for both effect and bias. (This is intended to be simple and not distracting, rather than optimal (e.g., see Table 3 in R 2023 Stat. Sci.).
- Using the 200 blocks with methylmercury levels (rather than all 406 blocks), rejection of no effect on HDL cholesterol level becomes sensitive at $\Gamma = 3.614$. No $\boldsymbol{\theta} \in B_{3.614}$ would lead to a P-value above 0.05. (Was $B_{4.5}$ for $I = 406$.)
- In parallel, using the same people in the same blocks, no $\boldsymbol{\theta} \in B_{1.993}$ is plausible if alcohol does not affect methylmercury levels, having been rejected in a 0.05 level test.

Unaffected Outcomes

- What should we make of the evidence from methylmercury of biased treatment assignment evident, $\boldsymbol{\theta} \neq \overline{\boldsymbol{\theta}}$, where $\overline{\theta}_{ij} = 1/J, \forall i,j$ and $B_1 = \{\overline{\boldsymbol{\theta}}\}$?
- To avoid many possible combinations, will do two-sided, 0.05-level Quade's tests for both effect and bias. (This is intended to be simple and not distracting, rather than optimal (e.g., see Table 3 in R 2023 Stat. Sci.).
- Using the 200 blocks with methylmercury levels (rather than all 406 blocks), rejection of no effect on HDL cholesterol level becomes sensitive at $\Gamma = 3.614$. No $\boldsymbol{\theta} \in B_{3.614}$ would lead to a P-value above 0.05. (Was $B_{4.5}$ for $I = 406$.)
- In parallel, using the same people in the same blocks, no $\boldsymbol{\theta} \in B_{1.993}$ is plausible if alcohol does not affect methylmercury levels, having been rejected in a 0.05 level test.
- The sensitivity analysis for HDL cholesterol doesn't require amendment, but it does leave us wondering about $\boldsymbol{\theta} \in B_{3.614} - B_{1.993}$; i.e., in $B_{3.614}$ but not in $B_{1.993}$.

Gaps Between Tests for Bias and Sensitivity Analyses

- Part of the boundary of $B_{3.614}$ is troublesome, because there is a $\boldsymbol{\theta} \in B_{3.615}$ that would lead us to accept no effect of alcohol on HDL cholesterol. Call these **troublesome boundary points** $\mathcal{J}$. What does methylmercury say about the troublesome boundary points $\mathcal{J}$?

Gaps Between Tests for Bias and Sensitivity Analyses

- Part of the boundary of $B_{3.614}$ is troublesome, because there is a $\boldsymbol{\theta} \in B_{3.615}$ that would lead us to accept no effect of alcohol on HDL cholesterol. Call these **troublesome boundary points** $\mathcal{J}$. What does methylmercury say about the troublesome boundary points $\mathcal{J}$?
- We would like to say: "no $\boldsymbol{\theta} \in \mathcal{J}$ is plausible." That would mean that the HDL cholesterol comparison isn't sensitive at $\Gamma = 3.614$ after all, but only to a larger Γ. If this were true, say that there is **no gap** between the test for bias using methylmercury and the sensitivity analysis for HDL cholesterol.

Gaps Between Tests for Bias and Sensitivity Analyses

- Part of the boundary of $B_{3.614}$ is troublesome, because there is a $\boldsymbol{\theta} \in B_{3.615}$ that would lead us to accept no effect of alcohol on HDL cholesterol. Call these **troublesome boundary points** $\mathcal{J}$. What does methylmercury say about the troublesome boundary points $\mathcal{J}$?
- We would like to say: "no $\boldsymbol{\theta} \in \mathcal{J}$ is plausible." That would mean that the HDL cholesterol comparison isn't sensitive at $\Gamma = 3.614$ after all, but only to a larger Γ. If this were true, say that there is **no gap** between the test for bias using methylmercury and the sensitivity analysis for HDL cholesterol.
- We can **test each of the troublesome** $\boldsymbol{\theta} \in \mathcal{J}$ using the methylmercury data. When we do this, the maximum P-value testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ for $\boldsymbol{\theta}_0 \in \mathcal{J}$ is 1.17×10^{-7}.

Gaps Between Tests for Bias and Sensitivity Analyses

- Part of the boundary of $B_{3.614}$ is troublesome, because there is a $\boldsymbol{\theta} \in B_{3.615}$ that would lead us to accept no effect of alcohol on HDL cholesterol. Call these **troublesome boundary points** $\mathcal{J}$. What does methylmercury say about the troublesome boundary points $\mathcal{J}$?
- We would like to say: "no $\boldsymbol{\theta} \in \mathcal{J}$ is plausible." That would mean that the HDL cholesterol comparison isn't sensitive at $\Gamma = 3.614$ after all, but only to a larger Γ. If this were true, say that there is **no gap** between the test for bias using methylmercury and the sensitivity analysis for HDL cholesterol.
- We can **test each of the troublesome** $\boldsymbol{\theta} \in \mathcal{J}$ using the methylmercury data. When we do this, the maximum P-value testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ for $\boldsymbol{\theta}_0 \in \mathcal{J}$ is 1.17×10^{-7}.
- The troublesome biases $\boldsymbol{\theta} \in \mathcal{J}$ are not plausible; so, there is **no gap**, and Γ must be larger than $\Gamma = 3.614$ to explain the higher HDL cholesterol levels of light daily drinkers. The evident bias in methylmercury **strengthened** the causal claim.

Confidence Set for $\boldsymbol{\theta}$; Informed Sensitivity Analyses

- What if we tested all $\boldsymbol{\theta}$'s? Let Θ be the set of all $\boldsymbol{\theta}_0$'s such that: (i) $1 = \sum_{j=1}^{J} \theta_{0ij}$, (ii) $0 \leq \theta_{0ij} \leq 1$, and (iii) the test using methylmercury does not reject $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ at the 0.05 level. An infinite set of IJ dimensional $\boldsymbol{\theta}_0$'s.

Confidence Set for $\boldsymbol{\theta}$; Informed Sensitivity Analyses

- What if we tested all $\boldsymbol{\theta}$'s? Let Θ be the set of all $\boldsymbol{\theta}_0$'s such that: (i) $1 = \sum_{j=1}^{J} \theta_{0ij}$, (ii) $0 \leq \theta_{0ij} \leq 1$, and (iii) the test using methylmercury does not reject $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ at the 0.05 level. An infinite set of IJ dimensional $\boldsymbol{\theta}_0$'s.
- A sensitivity analysis is **informed by a test for bias** if it is confined to $\boldsymbol{\theta} \in B_\Gamma \cap \Theta$:

$$P'_\Gamma = \max_{\boldsymbol{\theta} \in B_\Gamma \cap \Theta} \sum_{\mathbf{z} \in \mathcal{Z}} \left[\sum_i \varphi(w_i) \sum_j z_{ij}\, q_{ij} \geq t \right] \prod_i \prod_j \theta_{ij}^{z_{ij}}$$

instead of $\quad P_\Gamma = \max_{\boldsymbol{\theta} \in B_\Gamma} \sum_{\mathbf{z} \in \mathcal{Z}} \left[\sum_i \varphi(w_i) \sum_j z_{ij}\, q_{ij} \geq t \right] \prod_i \prod_j \theta_{ij}^{z_{ij}}$

Confidence Set for $\boldsymbol{\theta}$; Informed Sensitivity Analyses

- What if we tested all $\boldsymbol{\theta}$'s? Let Θ be the set of all $\boldsymbol{\theta}_0$'s such that: (i) $1 = \sum_{j=1}^{J} \theta_{0ij}$, (ii) $0 \leq \theta_{0ij} \leq 1$, and (iii) the test using methylmercury does not reject $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ at the 0.05 level. An infinite set of IJ dimensional $\boldsymbol{\theta}_0$'s.
- A sensitivity analysis is **informed by a test for bias** if it is confined to $\boldsymbol{\theta} \in B_\Gamma \cap \Theta$:

$$P'_\Gamma = \max_{\boldsymbol{\theta} \in B_\Gamma \cap \Theta} \sum_{\mathbf{z} \in \mathcal{Z}} \left[\sum_i \varphi(w_i) \sum_j z_{ij}\, q_{ij} \geq t \right] \prod_i \prod_j \theta_{ij}^{z_{ij}}$$

instead of $\quad P_\Gamma = \max_{\boldsymbol{\theta} \in B_\Gamma} \sum_{\mathbf{z} \in \mathcal{Z}} \left[\sum_i \varphi(w_i) \sum_j z_{ij}\, q_{ij} \geq t \right] \prod_i \prod_j \theta_{ij}^{z_{ij}}$

- Always, $P'_\Gamma \leq P_\Gamma$. For HDL cholesterol, $P_{3.614} = 0.05 = P'_{3.82}$.

Confidence Set for $\boldsymbol{\theta}$; Informed Sensitivity Analyses

- What if we tested all $\boldsymbol{\theta}$'s? Let Θ be the set of all $\boldsymbol{\theta}_0$'s such that: (i) $1 = \sum_{j=1}^{J} \theta_{0ij}$, (ii) $0 \leq \theta_{0ij} \leq 1$, and (iii) the test using methylmercury does not reject $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ at the 0.05 level. An infinite set of IJ dimensional $\boldsymbol{\theta}_0$'s.
- A sensitivity analysis is **informed by a test for bias** if it is confined to $\boldsymbol{\theta} \in B_\Gamma \cap \Theta$:

$$P'_\Gamma = \max_{\boldsymbol{\theta} \in B_\Gamma \cap \Theta} \sum_{\mathbf{z} \in \mathcal{Z}} \left[\sum_i \varphi(w_i) \sum_j z_{ij}\, q_{ij} \geq t \right] \prod_i \prod_j \theta_{ij}^{z_{ij}}$$

instead of $\quad P_\Gamma = \max_{\boldsymbol{\theta} \in B_\Gamma} \sum_{\mathbf{z} \in \mathcal{Z}} \left[\sum_i \varphi(w_i) \sum_j z_{ij}\, q_{ij} \geq t \right] \prod_i \prod_j \theta_{ij}^{z_{ij}}$

- Always, $P'_\Gamma \leq P_\Gamma$. For HDL cholesterol, $P_{3.614} = 0.05 = P'_{3.82}$.
- $\Gamma = 3.614$ is $(\Lambda,\ \Delta) = (6,\ 8.7)$, while $\Gamma = 3.82$ is $(\Lambda,\ \Delta) = (6,\ 10.1)$.

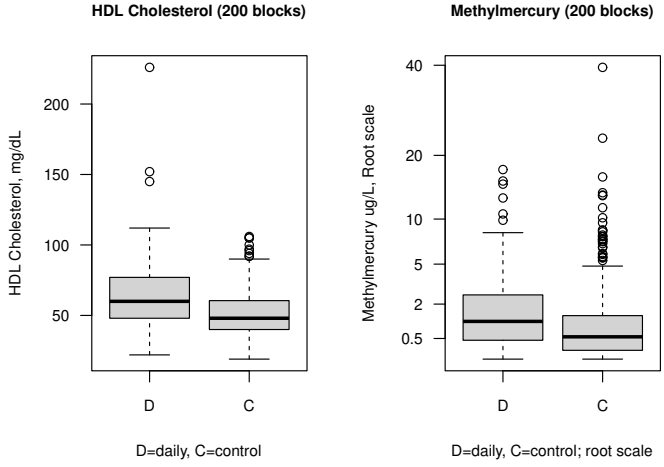

Figure: **Summary**: $\theta \neq \overline{\theta}$ from right. The smallest Γ explaining the right is too small to explain the left. The smallest Γ that explains both sides is larger than the smallest Γ that explains the left.

Summary

- Unmeasured biases are present in observational studies, but they may not be debilitating.

Summary

- Unmeasured biases are present in observational studies, but they may not be debilitating.
- Statistical theory can aid in avoiding mistakes that exaggerate the sensitivity to unmeasured biases. Large mistakes are possible in design and analysis.

Summary

- Unmeasured biases are present in observational studies, but they may not be debilitating.
- Statistical theory can aid in avoiding mistakes that exaggerate the sensitivity to unmeasured biases. Large mistakes are possible in design and analysis.
- Design sensitivity and Bahadur efficiency of a sensitivity analysis are two tools that guide design and analysis.

Summary

- Unmeasured biases are present in observational studies, but they may not be debilitating.
- Statistical theory can aid in avoiding mistakes that exaggerate the sensitivity to unmeasured biases. Large mistakes are possible in design and analysis.
- Design sensitivity and Bahadur efficiency of a sensitivity analysis are two tools that guide design and analysis.
- Evidence of biased treatment assignment may increase insensitivity to unmeasured bias.

Some references

- *Introduction to the Theory of Observational Studies*, Springer 2025.
- Bahadur efficiency of observational block designs. *JASA*, 2024;119:1871-1881.
- Can we reliably detect biases that matter in observational studies? *Statist. Sci.* 2023;38:440-457.
- Sensitivity analyses informed by tests for bias in observational studies. *Biometrics* 2023;79:475-487.
- A second evidence factor for a second control group. *Biometrics* 2023;79:3968-3980.
- Design sensitivity in observational studies. Biometrika 2004;91:153-164.
- *Design of Observational Studies*, 2^{nd} edition, New York: Springer, 2020.
- (with D.B. Rubin) Propensity scores in the design of observational studies for causal effects. Biometrika, 2023;110:1-13.

Understanding $\varphi(\cdot)$ in terms of abz(y) for $J = 2$

- For a single treated-minus-control matched pair difference, Y,

$$\text{abz}(y) = \Pr(Y > 0 | \, |Y| = y), \text{ for } y > 0,$$

from Albers, Bickel, van Zwet (1976, AOS, 4, 108-156).

Understanding $\varphi(\cdot)$ in terms of abz(y) for $J = 2$

- For a single treated-minus-control matched pair difference, Y,

$$\text{abz}(\gamma) = \Pr(Y > 0 | \, |Y| = \gamma), \text{ for } \gamma > 0,$$

  from Albers, Bickel, van Zwet (1976, AOS, 4, 108-156).
- Question: Suppose that you could observe an infinite number of pair differences, Y_i, but only for a single value of γ of $|Y| = \gamma$. What γ would you pick?

Understanding $\varphi(\cdot)$ in terms of abz(y) for $J = 2$

- For a single treated-minus-control matched pair difference, Y,

$$\text{abz}(\gamma) = \Pr(Y > 0 | \, |Y| = \gamma), \text{ for } \gamma > 0,$$

  from Albers, Bickel, van Zwet (1976, AOS, 4, 108-156).
- Question: Suppose that you could observe an infinite number of pair differences, Y_i, but only for a single value of γ of $|Y| = \gamma$. What γ would you pick?
- Given that $|Y| = \gamma$, a bias of Γ in the absence of a treatment effect cannot produce so many $Y > \gamma$ if

$$\text{abz}(\text{y}) > \frac{\Gamma}{\Gamma + 1}.$$

  So, the answer is the γ that maximizes $\text{abz}(\text{y})$. (Rosenbaum 2010 JASA 105, 692-702).

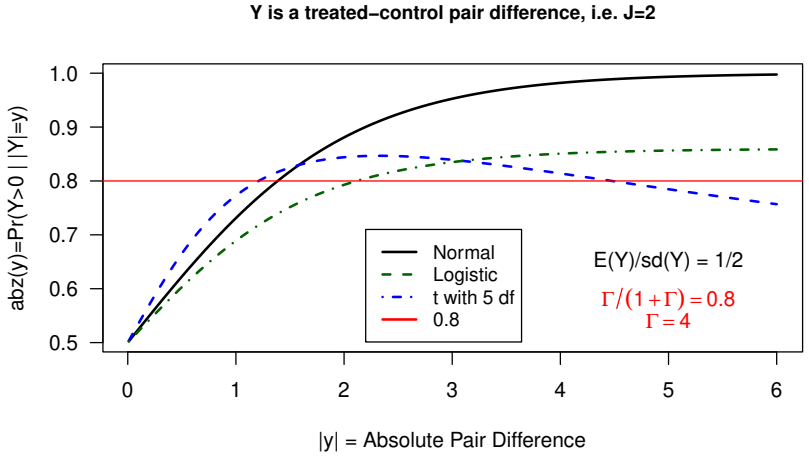

Figure: abz(y) plotted against $|y|$ where $Y - \delta$ is N(o,1), logistic, or t_5, and $\delta = \sigma/2$ so that $E(Y)/\sigma = 1/2$ for each distribution, where σ is the standard deviation of Y.

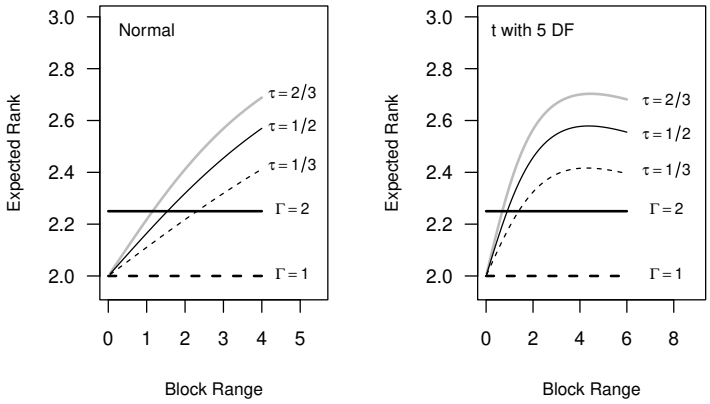

Figure: In 1-to-2 blocks of size $J = 3$, the curves show the expected within block rank — 1, 2, or 3 — conditionally given the within block range. Horizontal lines show maximum expectation with a bias of Γ and no treatment effect.

Adaptive Inference: Use Two φ's

- Do two tests with different φ's, and take the minimum of the two P-values as a test statistic, obtaining a P-value from it.
- Berk and Jones (1978) show this is "relatively optimal" in the sense of having the larger Bahadur efficiency of the two tests.
- This also works in sensitivity analyses, where it also has the larger of the two design sensitivities.

Berk, R.H., Jones, D.H. Relatively optimal combinations of test statistics. Scand J Stat 1978;5:158-62.
Rosenbaum, P.R. Testing one hypothesis twice … Biometrika 2012;99:763-74.
Rosenbaum, P.R. Bahadur efficiency of observational block designs. JASA 2024;119:1871-81.

Example of Adaptive Inference from `weightedRank` in R

```
library(weightedRank)
wgtRank(y,phi="wilc",gamma=4.5)
pval 0.7400862
wgtRank(y,phi="quade",gamma=4.5)
pval 0.04470762

Test-twice (tt) at gamma=4.4
wgtRanktt(y,phi1="wilc",phi2="quade",gamma=4.4)
jointP 0.04521279
cor12 0.8686031
Separate p-values
phi1 0.6719488
phi2 0.0312262
```

Proof of Ignorability Given ζ

- Essentially the same as the corresponding result for the propensity score $e(\mathbf{x})$.

Proof of Ignorability Given ζ

- Essentially the same as the corresponding result for the propensity score $e(\mathbf{x})$.
- **Must show**: $\Pr(Z = 1|\mathbf{x}, r_T, r_C) = \Pr(Z = 1|\zeta)$.

Proof of Ignorability Given ζ

- Essentially the same as the corresponding result for the propensity score $e(\mathbf{x})$.
- **Must show**: $\Pr(Z = 1|\mathbf{x}, r_T, r_C) = \Pr(Z = 1|\zeta)$.
- ζ is **by definition** $\zeta = \Pr(Z = 1|\mathbf{x}, r_T, r_C)$, so the task is to show $\zeta = \Pr(Z = 1|\zeta)$.

Proof of Ignorability Given ζ

- Essentially the same as the corresponding result for the propensity score $e(\mathbf{x})$.
- **Must show**: $\Pr(Z = 1|\mathbf{x}, r_T, r_C) = \Pr(Z = 1|\zeta)$.
- ζ is **by definition** $\zeta = \Pr(Z = 1|\mathbf{x}, r_T, r_C)$, so the task is to show $\zeta = \Pr(Z = 1|\zeta)$.
- Also, $\zeta = \Pr(Z = 1|\mathbf{x}, r_T, r_C)$ **is a function of** $(\mathbf{x}, r_T, r_C)$, so

$$\Pr(Z = 1|\mathbf{x}, r_T, r_C) = \Pr(Z = 1|\mathbf{x}, r_T, r_C, \zeta).$$

Proof of Ignorability Given ζ

- Essentially the same as the corresponding result for the propensity score $e(\mathbf{x})$.
- **Must show**: $\Pr(Z = 1|\mathbf{x}, r_T, r_C) = \Pr(Z = 1|\zeta)$.
- ζ is **by definition** $\zeta = \Pr(Z = 1|\mathbf{x}, r_T, r_C)$, so the task is to show $\zeta = \Pr(Z = 1|\zeta)$.
- Also, $\zeta = \Pr(Z = 1|\mathbf{x}, r_T, r_C)$ **is a function of** $(\mathbf{x}, r_T, r_C)$, so

$$\Pr(Z = 1|\mathbf{x}, r_T, r_C) = \Pr(Z = 1|\mathbf{x}, r_T, r_C, \zeta).$$

- Trivially,

$$\Pr(Z = 1|\zeta) = \mathrm{E}\{\Pr(Z = 1|\mathbf{x}, r_T, r_C, \zeta) \,|\, \zeta\}$$
$$= \mathrm{E}\{\Pr(Z = 1|\mathbf{x}, r_T, r_C) \,|\, \zeta\} = \mathrm{E}(\zeta \,|\, \zeta) = \zeta,$$

  as required to complete the proof.

R Code from `weightedRank`

```
ef2C(hdl3,gamma=4,upsilon = 3.75)$pvals
```

TreatedVSControl1 0.11069568
Control2vsOthers 0.11173143
Combined 0.04667447

```
p1=dwgtRank(hdl3[,1:2],gamma=4,m=8, m1=7, m2=8)$pval
```

```
p2=dwgtRank(hdl3[,3:1],gamma=3.75,alternative="less",m=8,m1=8,m2=8,
range=FALSE,scores=c(1,2,5))$pval
```

```
c(p1,p2)
```

0.1106957 0.1117314

```
sensitivitymv::truncatedP(c(p1,p2))
```

0.04667447

Bahadur Efficiencies for Pairs, $J = 2$

Table: Efficiency of a sensitivity analysis at Γ **vs. U868** with Normal errors and $\tau = 1/2$ of the standard deviation of a treated-minus-control pair difference. The best result is in **bold**.

		Normal Errors, Paired Data, $J = 2$			
		Wilcoxon	Quade	U868	U878
Γ	$\widetilde{\Gamma}$	2.2	3.2	4.2	**5.1**
1		0.72	**1.05**	1.00	0.92
1.5		0.36	0.86	**1.00**	**1.00**
2		0.06	0.63	1.00	**1.11**
3		0.00	0.05	1.00	**1.70**
4		0.00	0.00	1.00	**15.56**

- By definition, efficiency of U868 is 1.00.
- Quade=Wilcoxon's signed rank best at $\Gamma = 1$, but not at $\Gamma = 1.5$.

Bahadur Efficiencies for 1-to-3 Blocks, $J = 4$

Table: Efficiency **relative to U868** with Normal errors and $\tau = 1/2$ of the standard deviation of a treated-minus-control pair difference. The best result is in **bold**.

		Normal Errors, 1-to-3 Blocks, $J = 4$			
		Wilcoxon	Quade	U868	U878
Γ	$\widetilde{\Gamma}$	3.5	4.4	5.2	**5.7**
1		1.08	**1.21**	1.00	0.85
1.5		0.83	**1.11**	1.00	0.89
2		0.58	**1.01**	1.00	0.93
3		0.15	0.76	1.00	**1.04**
4		0.00	0.23	1.00	**1.41**

- Quade's statistic does well for $\Gamma \leq 2$ but falls behind for $\Gamma \geq 3$.

$J > 2$ Needs a Larger Γ Than $J = 2$ to Produce the Same Mean

- In the favorable situation, Treated are N(1/2, 1), Controls are N(0, 1) in $I = 100,000$ blocks of size $J = 4$.
- In the unfavorable situations, order statistics from the one treated individual and $J - 1$ of the controls are reallocated to "treatment" or "control" under the sensitivity model with θ_{ij} that maximize the "treated" group's expectation for $\Gamma = 2.5$.

Situation	T-mean	C-mean	Difference	T-sd	C-sd	Stand-diff
Favorable, $J = 4$	0.50	0.00	0.50	1.00	1.00	0.35
Unfavorable, $J = 4$	0.45	0.02	0.43	1.02	1.00	0.30
Unfavorable, $J = 2$	0.50	0.00	0.51	1.00	1.00	0.36

Randomized Trials of Alcohol and HDL-C

- Haskell, W.L. et al, 1984. The effect of cessation and resumption of moderate alcohol intake on serum high-density-lipoprotein subfractions: A controlled study. *New England Journal of Medicine*, 310(13), pp.805-810.
- Burr, M.L. et. al, 1986. Alcohol and high-density-lipoprotein cholesterol: A randomized controlled trial. *British Journal of Nutrition*, 56(1), pp.81-86.
- Gepner, Y. et al, 2015. Effects of initiating moderate alcohol intake on cardiometabolic risk in adults with type 2 diabetes: a 2-year randomized, controlled trial. *Annals of Internal Medicine*, 163(8), pp.569-579.