Lord’s Paradox and Consequences for Effects of Interventions on Outcomes

Perman Gochyyev and Mark Wilson

BEAR Center, Graduate School of Education
University of California, Berkeley

February 1, 2022

outline

1. Lord’s paradox and real examples (15 mins)
2. A closer look (20 mins)
3. Which approach should I use? (10 mins)

outline

1. Lord's paradox and real examples (15 mins)
2. A closer look (20 mins)
3. Which approach should I use? (10 mins)

outline

1. Lord's paradox and real examples (15 mins)
2. A closer look (20 mins)
3. Which approach should I use? (10 mins)

Paradox

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin, 68*(5), 304-305.

on paradox

- Pearl (2014): *"Among the many peculiarities that were dubbed "paradoxes" by well meaning statisticians, the one reported by Frederic M. Lord in 1967 has earned a special status."*
  - *"Unlike Simpson's reversal, Lord's is easier to state, harder to disentangle"*
  - *"... and, for some reason, it has been lingering for almost four decades, under several interpretations and re-interpretations, and it keeps coming up in new situations and under new lights"*
  - *"... the original version presented by Lord, to the best of my knowledge, has not been given a proper treatment, not to mention a resolution"*

on paradox

- Pearl (2014): *"Among the many peculiarities that were dubbed "paradoxes" by well meaning statisticians, the one reported by Frederic M. Lord in 1967 has earned a special status."*
  - *"Unlike Simpson's reversal, Lord's is easier to state, harder to disentangle"*
  - *"... and, for some reason, it has been lingering for almost four decades, under several interpretations and re-interpretations, and it keeps coming up in new situations and under new lights"*
  - *"... the original version presented by Lord, to the best of my knowledge, has not been given a proper treatment, not to mention a resolution"*

on paradox

- Pearl (2014): *"Among the many peculiarities that were dubbed "paradoxes" by well meaning statisticians, the one reported by Frederic M. Lord in 1967 has earned a special status."*
  - *"Unlike Simpson's reversal, Lord's is easier to state, harder to disentangle"*
  - *"... and, for some reason, it has been lingering for almost four decades, under several interpretations and re-interpretations, and it keeps coming up in new situations and under new lights"*
  - *"... the original version presented by Lord, to the best of my knowledge, has not been given a proper treatment, not to mention a resolution"*

on paradox

- Pearl (2014): *"Among the many peculiarities that were dubbed "paradoxes" by well meaning statisticians, the one reported by Frederic M. Lord in 1967 has earned a special status."*
    - *"Unlike Simpson's reversal, Lord's is easier to state, harder to disentangle"*
    - *"... and, for some reason, it has been lingering for almost four decades, under several interpretations and re-interpretations, and it keeps coming up in new situations and under new lights"*
    - *"... the original version presented by Lord, to the best of my knowledge, has not been given a proper treatment, not to mention a resolution"*

Lord, 1967: original context

- Two groups: boys and girls
- weight in September (Pretest)
- weight in June (Posttest)

Lord, 1967: original context

- Two groups: boys and girls
- weight in September (Pretest)
- weight in June (Posttest)

Lord, 1967: original context

- Two groups: boys and girls
- weight in September (Pretest)
- weight in June (Posttest)

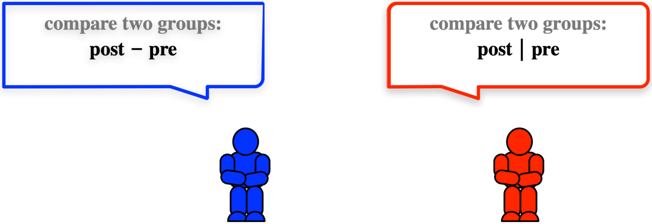
Two statisticians

Data

Statistician 1

Statistician 2

Two statisticians



compare two groups:
post − pre

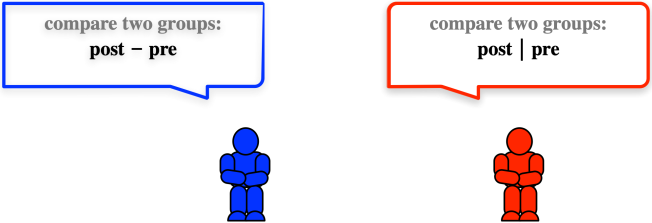
compare two groups:
post | pre

Statistician 1

Statistician 2

Two statisticians

CS approach



compare two groups:
post − pre

compare two groups:
post | pre

Statistician 1

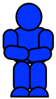
Statistician 2

Two statisticians

CS approach

compare two groups:

post − pre

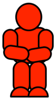


Statistician 1

RV approach

compare two groups:

post | pre



Statistician 2

Statistician 1

- What is:

$$(Weight_{June} - Weight_{Sept})^{girls} - (Weight_{June} - Weight_{Sept})^{boys}$$
$$?$$

- To put it in a more formal equation:
  Girls (G=1), Boys (G=0)

$$W_{post} - W_{pre} = \beta_1 + \beta_2 G + \epsilon$$

Statistician 1

- What is:

$$(Weight_{June} - Weight_{Sept})^{girls} - (Weight_{June} - Weight_{Sept})^{boys}$$

?

- To put it in a more formal equation:
  Girls (G=1), Boys (G=0)

$$W_{post} - W_{pre} = \beta_1 + \boldsymbol{\beta_2} G + \epsilon$$

Statistician 2:

- Uses RV approach, and covaries pretest out of posttest:

$$(Weight^{girls}_{June} - Weight^{boys}_{June}) \Big| Weight_{Sept}$$

- or, more formally:

$$W_{post} = \beta_1 + \boldsymbol{\beta_2} G + \beta_3 W_{pre} + \epsilon$$

Statistician 2:

- Uses RV approach, and covaries pretest out of posttest:

$$\left(Weight^{girls}_{June} - Weight^{boys}_{June}\right) \Big| \, Weight_{Sept}$$

- or, more formally:

$$W_{post} = \beta_1 + \boldsymbol{\beta_2} G + \beta_3 W_{pre} + \epsilon$$

summaries from the data

Data

- $mean[Weight^{girls}_{Sept}] = mean[Weight^{girls}_{June}]$
- $mean[Weight^{boys}_{Sept}] = mean[Weight^{boys}_{June}]$

summaries from the data

Data

- $mean[Weight^{girls}_{Sept}] = mean[Weight^{girls}_{June}]$

- $mean[Weight^{boys}_{Sept}] = mean[Weight^{boys}_{June}]$

Two different conclusions

Statistician 1

- For girls: no difference between weights in September and June
- For boys: no difference between weights in September and June
- Conclusion: no differences between boys and girls in weight gain.

Statistician 2

- "... boys showed significantly more gain in weight than the girls when proper allowance is made for differences in initial weight between the two sexes"
- Adjusting/controlling for weight in September, boys are higher in their weight in June.
- Conclusion: boys showed more gain than girls.

Two different conclusions

Statistician 1

- For girls: no difference between weights in September and June
- For boys: no difference between weights in September and June
- Conclusion: no differences between boys and girls in weight gain.

Statistician 2

- "... boys showed significantly more gain in weight than the girls when proper allowance is made for differences in initial weight between the two sexes"
- Adjusting/controlling for weight in September, boys are higher in their weight in June.
- Conclusion: boys showed more gain than girls.

Two different conclusions

Statistician 1

- For girls: no difference between weights in September and June
- For boys: no difference between weights in September and June
- Conclusion: no differences between boys and girls in weight gain.

Statistician 2

- "... boys showed significantly more gain in weight than the girls when proper allowance is made for differences in initial weight between the two sexes"
- Adjusting/controlling for weight in September, boys are higher in their weight in June.
- Conclusion: boys showed more gain than girls.

Two different conclusions

Statistician 1

- For girls: no difference between weights in September and June
- For boys: no difference between weights in September and June
- Conclusion: no differences between boys and girls in weight gain.

Statistician 2

- "... boys showed significantly more gain in weight than the girls when proper allowance is made for differences in initial weight between the two sexes"
- Adjusting/controlling for weight in September, boys are higher in their weight in June.
- Conclusion: boys showed more gain than girls.

Two different conclusions

Statistician 1

- For girls: no difference between weights in September and June
- For boys: no difference between weights in September and June
- Conclusion: no differences between boys and girls in weight gain.

Statistician 2

- "... boys showed significantly more gain in weight than the girls when proper allowance is made for differences in initial weight between the two sexes"
- Adjusting/controlling for weight in September, boys are higher in their weight in June.
- Conclusion: boys showed more gain than girls.

Two different conclusions

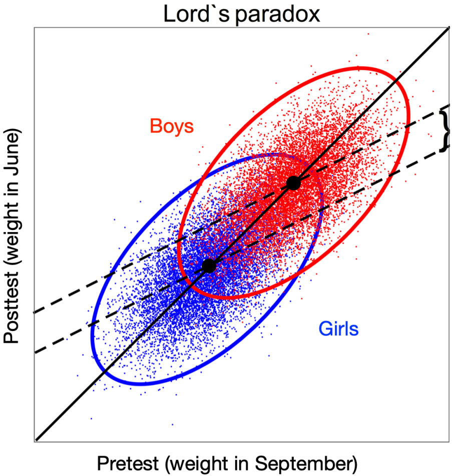
Statistician 1

- For girls: no difference between weights in September and June
- For boys: no difference between weights in September and June
- Conclusion: no differences between boys and girls in weight gain.

Statistician 2

- "... boys showed significantly more gain in weight than the girls when proper allowance is made for differences in initial weight between the two sexes"
- Adjusting/controlling for weight in September, boys are higher in their weight in June.
- Conclusion: boys showed more gain than girls.

Lord’s paradox

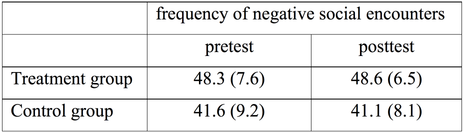


Real example 1 (Allison, 1990)

	frequency of negative social encounters	
	pretest	posttest
Treatment group	48.3 (7.6)	48.6 (6.5)
Control group	41.6 (9.2)	41.1 (8.1)

- Treatment: plastic surgery *(n=18)*
- Treatment group: children with craniofacial abnormalities *(n=18)*
- Control group: children (with same age range) without craniofacial abnormalities *(n=30)*
- Outcome: frequency of negative social encounters (based on parental reports)
- Statistician 1: treatment effect $= 0$
- Statistician 2: treatment effect> 0, significant at p-value$= 0.03$
    - Conclusion: plastic surgery had a negative effect on children's social experiences

Real example 1 (Allison, 1990)



	frequency of negative social encounters	
	pretest	posttest
Treatment group	48.3 (7.6)	48.6 (6.5)
Control group	41.6 (9.2)	41.1 (8.1)

- Treatment: plastic surgery *(n=18)*
- Treatment group: children with craniofacial abnormalities *(n=18)*
- Control group: children (with same age range) without craniofacial abnormalities *(n=30)*
- Outcome: frequency of negative social encounters (based on parental reports)
- Statistician 1: treatment effect $= 0$
- Statistician 2: treatment effect> 0, significant at p-value$= 0.03$
  - Conclusion: plastic surgery had a negative effect on children's social experiences

Real example 1 (Allison, 1990)

	frequency of negative social encounters	
	pretest	posttest
Treatment group	48.3 (7.6)	48.6 (6.5)
Control group	41.6 (9.2)	41.1 (8.1)

- Treatment: plastic surgery *(n=18)*
- Treatment group: children with craniofacial abnormalities *(n=18)*
- Control group: children (with same age range) without craniofacial abnormalities *(n=30)*
- Outcome: frequency of negative social encounters (based on parental reports)

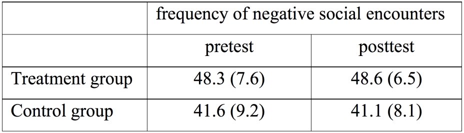
- Statistician 1: treatment effect $= 0$
- Statistician 2: treatment effect> 0, significant at p-value$= 0.03$
  - Conclusion: plastic surgery had a negative effect on children's social experiences

Real example 1 (Allison, 1990)

	frequency of negative social encounters	
	pretest	posttest
Treatment group	48.3 (7.6)	48.6 (6.5)
Control group	41.6 (9.2)	41.1 (8.1)

- Treatment: plastic surgery *(n=18)*
- Treatment group: children with craniofacial abnormalities *(n=18)*
- Control group: children (with same age range) without craniofacial abnormalities *(n=30)*
- Outcome: frequency of negative social encounters (based on parental reports)
- Statistician 1: treatment effect $= 0$
- Statistician 2: treatment effect> 0, significant at p-value$= 0.03$
  - Conclusion: plastic surgery had a negative effect on children's social experiences

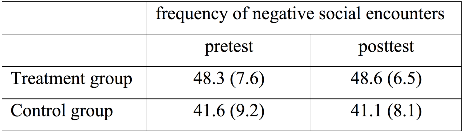
Real example 1 (Allison, 1990)



	frequency of negative social encounters	
	pretest	posttest
Treatment group	48.3 (7.6)	48.6 (6.5)
Control group	41.6 (9.2)	41.1 (8.1)

- Treatment: plastic surgery *(n=18)*
- Treatment group: children with craniofacial abnormalities *(n=18)*
- Control group: children (with same age range) without craniofacial abnormalities *(n=30)*
- Outcome: frequency of negative social encounters (based on parental reports)
- Statistician 1: treatment effect = 0
- Statistician 2: treatment effect> 0, significant at p-value= 0.03
  - Conclusion: plastic surgery had a negative effect on children's social experiences

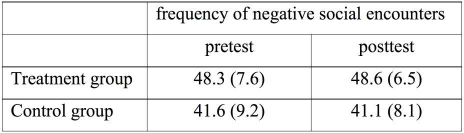
Real example 1 (Allison, 1990)



	frequency of negative social encounters	
	pretest	posttest
Treatment group	48.3 (7.6)	48.6 (6.5)
Control group	41.6 (9.2)	41.1 (8.1)

- Treatment: plastic surgery *(n=18)*
- Treatment group: children with craniofacial abnormalities *(n=18)*
- Control group: children (with same age range) without craniofacial abnormalities *(n=30)*
- Outcome: frequency of negative social encounters (based on parental reports)
- Statistician 1: treatment effect $= 0$
- Statistician 2: treatment effect> 0, significant at p-value$= 0.03$
  - Conclusion: plastic surgery had a negative effect on children's social experiences

Real example 1 (Allison, 1990)



	frequency of negative social encounters	
	pretest	posttest
Treatment group	48.3 (7.6)	48.6 (6.5)
Control group	41.6 (9.2)	41.1 (8.1)

- Treatment: plastic surgery *(n=18)*
- Treatment group: children with craniofacial abnormalities *(n=18)*
- Control group: children (with same age range) without craniofacial abnormalities *(n=30)*
- Outcome: frequency of negative social encounters (based on parental reports)
- Statistician 1: treatment effect $= 0$
- Statistician 2: treatment effect> 0, significant at p-value$= 0.03$
  - Conclusion: plastic surgery had a negative effect on children's social experiences

Real example 2 (Theobald, 2015)

- Washington State public schools (Theobald, 2015)
- Treatment: Special education services
- Treatment group: students who were not already receiving special education services at the beginning of the school year, and placed into a Specific Learning Disability category
- Control group: all other students
- Outcome: reading and math scores

Real example 2 (Theobald, 2015)

- Washington State public schools (Theobald, 2015)
- Treatment: Special education services
- Treatment group: students who were not already receiving special education services at the beginning of the school year, and placed into a Specific Learning Disability category
- Control group: all other students
- Outcome: reading and math scores

Real example 2 (Theobald, 2015)

- Washington State public schools (Theobald, 2015)
- Treatment: Special education services
- Treatment group: students who were not already receiving special education services at the beginning of the school year, and placed into a Specific Learning Disability category
- Control group: all other students
- Outcome: reading and math scores

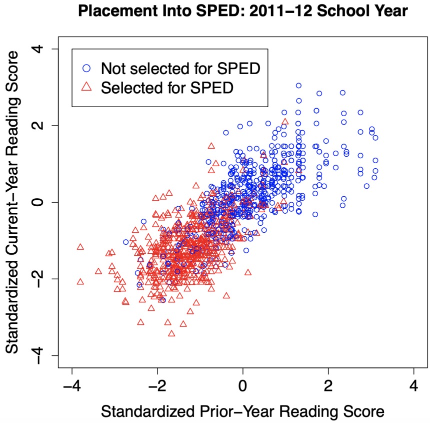
Real example 2 (Theobald, 2015)

- Washington State public schools (Theobald, 2015)
- Treatment: Special education services
- Treatment group: students who were not already receiving special education services at the beginning of the school year, and placed into a Specific Learning Disability category
- Control group: all other students
- Outcome: reading and math scores

Real example 2 (Theobald, 2015)

- Washington State public schools (Theobald, 2015)
- Treatment: Special education services
- Treatment group: students who were not already receiving special education services at the beginning of the school year, and placed into a Specific Learning Disability category
- Control group: all other students
- Outcome: reading and math scores

Real example 2



Real example 2 (Theobald, 2015)

- Statistician 1: special education services have no impact on student test performance
- Statistician 2: special education services have a large negative impact on student test performance

Real example 2 (Theobald, 2015)

- Statistician 1: special education services have no impact on student test performance
- Statistician 2: special education services have a large negative impact on student test performance

Paradox

This is Lord's paradox.

Paradox

Why does the approach taken by Statistician 2 — the method that currently dominates social science methodology — give an unintuitive and misleading result?

in other disciplines and traditions

- Causal SEM framework
  - Statistician 1: total effect
  - Statistician 2: direct effect (adjusting for pretest)
- Econometrics
  - Statistician 1: DID (also FD approach, same as FE estimator when two time-points)
  - Statistician 2: lagged dependent variable approach
- Experimental design
  - Statistician 1: ANOVA (RANOVA: repeated measures ANOVA)
  - Statistician 2: ANCOVA (analysis of covariance)

in other disciplines and traditions

- Causal SEM framework
  - Statistician 1: total effect
  - Statistician 2: direct effect (adjusting for pretest)
- Econometrics
  - Statistician 1: DID (also FD approach, same as FE estimator when two time-points)
  - Statistician 2: lagged dependent variable approach
- Experimental design
  - Statistician 1: ANOVA (RANOVA: repeated measures ANOVA)
  - Statistician 2: ANCOVA (analysis of covariance)

in other disciplines and traditions

- Causal SEM framework
  - Statistician 1: total effect
  - Statistician 2: direct effect (adjusting for pretest)
- Econometrics
  - Statistician 1: DID (also FD approach, same as FE estimator when two time-points)
  - Statistician 2: lagged dependent variable approach
- Experimental design
  - Statistician 1: ANOVA (RANOVA: repeated measures ANOVA)
  - Statistician 2: ANCOVA (analysis of covariance)

in other disciplines and traditions

- Causal SEM framework
  - Statistician 1: total effect
  - Statistician 2: direct effect (adjusting for pretest)
- Econometrics
  - Statistician 1: DID (also FD approach, same as FE estimator when two time-points)
  - Statistician 2: lagged dependent variable approach
- Experimental design
  - Statistician 1: ANOVA (RANOVA: repeated measures ANOVA)
  - Statistician 2: ANCOVA (analysis of covariance)

in other disciplines and traditions

- Causal SEM framework
  - Statistician 1: total effect
  - Statistician 2: direct effect (adjusting for pretest)
- Econometrics
  - Statistician 1: DID (also FD approach, same as FE estimator when two time-points)
  - Statistician 2: lagged dependent variable approach
- Experimental design
  - Statistician 1: ANOVA (RANOVA: repeated measures ANOVA)
  - Statistician 2: ANCOVA (analysis of covariance)

in other disciplines and traditions

- Causal SEM framework
  - Statistician 1: total effect
  - Statistician 2: direct effect (adjusting for pretest)
- Econometrics
  - Statistician 1: DID (also FD approach, same as FE estimator when two time-points)
  - Statistician 2: lagged dependent variable approach
- Experimental design
  - Statistician 1: ANOVA (RANOVA: repeated measures ANOVA)
  - Statistician 2: ANCOVA (analysis of covariance)

in other disciplines and traditions

- Causal SEM framework
  - Statistician 1: total effect
  - Statistician 2: direct effect (adjusting for pretest)
- Econometrics
  - Statistician 1: DID (also FD approach, same as FE estimator when two time-points)
  - Statistician 2: lagged dependent variable approach
- Experimental design
  - Statistician 1: ANOVA (RANOVA: repeated measures ANOVA)
  - Statistician 2: ANCOVA (analysis of covariance)

in other disciplines and traditions

- Causal SEM framework
  - Statistician 1: total effect
  - Statistician 2: direct effect (adjusting for pretest)
- Econometrics
  - Statistician 1: DID (also FD approach, same as FE estimator when two time-points)
  - Statistician 2: lagged dependent variable approach
- Experimental design
  - Statistician 1: ANOVA (RANOVA: repeated measures ANOVA)
  - Statistician 2: ANCOVA (analysis of covariance)

in other disciplines and traditions

- Causal SEM framework
  - Statistician 1: total effect
  - Statistician 2: direct effect (adjusting for pretest)
- Econometrics
  - Statistician 1: DID (also FD approach, same as FE estimator when two time-points)
  - Statistician 2: lagged dependent variable approach
- Experimental design
  - Statistician 1: ANOVA (RANOVA: repeated measures ANOVA)
  - Statistician 2: ANCOVA (analysis of covariance)

on paradox

- Senn (2008): *"In a disturbing paper in the Psychological Bulletin in 1967, Lord considered a case ..."*
- Rubin, Stuart & Zanutto (2003): *"A Classic Example of Poorly Formulated Causal Assessment—Lord's paradox"*
- Wainer & Brown (2007): *"..by far, the most difficult paradox to disentangle and requires clear thinking"*
- Lord (1967):
  "... there are as many different explanations as there are explainers"
  "... there simply is no logical or statistical procedure that can be counted on to make proper allowances for uncontrolled preexisting differences between groups"

on paradox

- Senn (2008): *"In a disturbing paper in the Psychological Bulletin in 1967, Lord considered a case ..."*
- Rubin, Stuart & Zanutto (2003): *"A Classic Example of Poorly Formulated Causal Assessment—Lord's paradox"*
- Wainer & Brown (2007): *"..by far, the most difficult paradox to disentangle and requires clear thinking"*
- Lord (1967):
  "... there are as many different explanations as there are explainers"
  "... there simply is no logical or statistical procedure that can be counted on to make proper allowances for uncontrolled preexisting differences between groups"

on paradox

- Senn (2008): *"In a disturbing paper in the Psychological Bulletin in 1967, Lord considered a case ..."*
- Rubin, Stuart & Zanutto (2003): *"A Classic Example of Poorly Formulated Causal Assessment—Lord's paradox"*
- Wainer & Brown (2007): *"..by far, the most difficult paradox to disentangle and requires clear thinking"*
- Lord (1967):
  "... there are as many different explanations as there are explainers"
  "... there simply is no logical or statistical procedure that can be counted on to make proper allowances for uncontrolled preexisting differences between groups"

on paradox

- Senn (2008): *"In a disturbing paper in the Psychological Bulletin in 1967, Lord considered a case ..."*
- Rubin, Stuart & Zanutto (2003): *"A Classic Example of Poorly Formulated Causal Assessment—Lord's paradox"*
- Wainer & Brown (2007): *"..by far, the most difficult paradox to disentangle and requires clear thinking"*
- Lord (1967):
  "... there are as many different explanations as there are explainers"
  "... there simply is no logical or statistical procedure that can be counted on to make proper allowances for uncontrolled preexisting differences between groups"

Three camps

- Debates over this paradox spread into mainly three directions:
  - Education/psychology methodologists: debates over reliability, measurement error, regression to the mean (Cronbach & Furby, 1970; Linn & Slinde, 1977 vs. Rogosa & Willett, 1983; Zimmerman & Williams, 1982)
  - Whose causal framework is better armed to explain the paradox (Holland & Rubin, 1982; Wainer & Brown, 2007; Pearl, 2014)
  - "It depends" camp (Kenny, 1975, Allison, 1990)

Three camps

- Debates over this paradox spread into mainly three directions:
  - Education/psychology methodologists: debates over reliability, measurement error, regression to the mean (Cronbach & Furby, 1970; Linn & Slinde, 1977 vs. Rogosa & Willett, 1983; Zimmerman & Williams, 1982)
  - Whose causal framework is better armed to explain the paradox (Holland & Rubin, 1982; Wainer & Brown, 2007; Pearl, 2014)
  - "It depends" camp (Kenny, 1975, Allison, 1990)

Three camps

- Debates over this paradox spread into mainly three directions:
  - Education/psychology methodologists: debates over reliability, measurement error, regression to the mean
    (Cronbach & Furby, 1970; Linn & Slinde, 1977 vs. Rogosa & Willett, 1983; Zimmerman & Williams, 1982)
  - Whose causal framework is better armed to explain the paradox
    (Holland & Rubin, 1982; Wainer & Brown, 2007; Pearl, 2014)
  - "It depends" camp
    (Kenny, 1975, Allison, 1990)

Three camps

- Debates over this paradox spread into mainly three directions:
    - Education/psychology methodologists: debates over reliability, measurement error, regression to the mean (Cronbach & Furby, 1970; Linn & Slinde, 1977 vs. Rogosa & Willett, 1983; Zimmerman & Williams, 1982)
    - Whose causal framework is better armed to explain the paradox (Holland & Rubin, 1982; Wainer & Brown, 2007; Pearl, 2014)
    - "It depends" camp (Kenny, 1975, Allison, 1990)

Educational researchers on change scores

- Cronbach & Furby (1970): *"It appears that investigators who ask questions regarding gain scores would ordinarily be better advised to frame their questions in other ways."*
- Linn & Slinde (1977): *"Problems in measuring change abound and the virtues in doing so are hard to find."*

Educational researchers on change scores

- Cronbach & Furby (1970): *"It appears that investigators who ask questions regarding gain scores would ordinarily be better advised to frame their questions in other ways."*
- Linn & Slinde (1977): *"Problems in measuring change abound and the virtues in doing so are hard to find."*

Educational researchers on change scores

- Main "issues":
  - "Change scores will be higher for individuals with a lower pretest" ("unfairness")
    - Counterexample in education: those who are high in the initial status might be better suited to understand the new instruction and gain more than those with a lower initial status
  - "Unreliable" (Gulliksen's 1950 formula)
    - Reliability (or unreliability) of scores relates to individual-level changes, but evaluation of the treatment effect is at the group level.
    - O'Brian (1998) provides a typical scenario, in which, for the group with N=25 persons, the aggregate-level reliability is 0.93 when it is only 0.33 at the individual level.

Educational researchers on change scores

- Main "issues":
  - "Change scores will be higher for individuals with a lower pretest" ("unfairness")
    - Counterexample in education: those who are high in the initial status might be better suited to understand the new instruction and gain more than those with a lower initial status
  - "Unreliable" (Gulliksen's 1950 formula)
    - Reliability (or unreliability) of scores relates to individual-level changes, but evaluation of the treatment effect is at the group level.
    - O'Brian (1998) provides a typical scenario, in which, for the group with N=25 persons, the aggregate-level reliability is 0.93 when it is only 0.33 at the individual level.

Educational researchers on change scores

- Main "issues":
  - "Change scores will be higher for individuals with a lower pretest" ("unfairness")
    - Counterexample in education: those who are high in the initial status might be better suited to understand the new instruction and gain more than those with a lower initial status
  - "Unreliable" (Gulliksen's 1950 formula)
    - Reliability (or unreliability) of scores relates to individual-level changes, but evaluation of the treatment effect is at the group level.
    - O'Brian (1998) provides a typical scenario, in which, for the group with N=25 persons, the aggregate-level reliability is 0.93 when it is only 0.33 at the individual level.

Educational researchers on change scores

- Main "issues":
  - "Change scores will be higher for individuals with a lower pretest" ("unfairness")
    - Counterexample in education: those who are high in the initial status might be better suited to understand the new instruction and gain more than those with a lower initial status
  - "Unreliable" (Gulliksen's 1950 formula)
    - Reliability (or unreliability) of scores relates to individual-level changes, but evaluation of the treatment effect is at the group level.
    - O'Brian (1998) provides a typical scenario, in which, for the group with N=25 persons, the aggregate-level reliability is 0.93 when it is only 0.33 at the individual level.

Educational researchers on change scores

- Main "issues":
  - "Change scores will be higher for individuals with a lower pretest" ("unfairness")
    - Counterexample in education: those who are high in the initial status might be better suited to understand the new instruction and gain more than those with a lower initial status
  - "Unreliable" (Gulliksen's 1950 formula)
    - Reliability (or unreliability) of scores relates to individual-level changes, but evaluation of the treatment effect is at the group level.
    - O'Brian (1998) provides a typical scenario, in which, for the group with N=25 persons, the aggregate-level reliability is 0.93 when it is only 0.33 at the individual level.

Educational researchers on change scores

- Main "issues":
  - "Change scores will be higher for individuals with a lower pretest" ("unfairness")
    - Counterexample in education: those who are high in the initial status might be better suited to understand the new instruction and gain more than those with a lower initial status
  - "Unreliable" (Gulliksen's 1950 formula)
    - Reliability (or unreliability) of scores relates to individual-level changes, but evaluation of the treatment effect is at the group level.
    - O'Brian (1998) provides a typical scenario, in which, for the group with N=25 persons, the aggregate-level reliability is 0.93 when it is only 0.33 at the individual level.

Neyman-Rubin framework

- Rubin, et al., (2003); Holland & Rubin (1983):
  - Suggest that Lord's example was a "poorly formulated causal assessment" since the potential outcome under the control diet is missing
  - *"...researcher investigating gain wouldn't know if changes in scores would have occurred with no treatment anyway"*
- However, the hypothetical researcher in Lord (1967) is interested in gender differences and not in the effect of the diet
  - Lord (1967): "differential effect" of the diet
- under the Neyman-Rubin (a.k.a. potential outcomes) causal framework the effect of gender cannot be a causal research question

Neyman-Rubin framework

- Rubin, et al., (2003); Holland & Rubin (1983):
    - Suggest that Lord's example was a "poorly formulated causal assessment" since the potential outcome under the control diet is missing
    - *"...researcher investigating gain wouldn't know if changes in scores would have occurred with no treatment anyway"*
- However, the hypothetical researcher in Lord (1967) is interested in gender differences and not in the effect of the diet
    - Lord (1967): "differential effect" of the diet
- under the Neyman-Rubin (a.k.a. potential outcomes) causal framework the effect of gender cannot be a causal research question

Neyman-Rubin framework

- Rubin, et al., (2003); Holland & Rubin (1983):
  - Suggest that Lord's example was a "poorly formulated causal assessment" since the potential outcome under the control diet is missing
  - *"...researcher investigating gain wouldn't know if changes in scores would have occurred with no treatment anyway"*
- However, the hypothetical researcher in Lord (1967) is interested in gender differences and not in the effect of the diet
  - Lord (1967): "differential effect" of the diet
- under the Neyman-Rubin (a.k.a. potential outcomes) causal framework the effect of gender cannot be a causal research question

Neyman-Rubin framework

- Rubin, et al., (2003); Holland & Rubin (1983):
  - Suggest that Lord's example was a "poorly formulated causal assessment" since the potential outcome under the control diet is missing
  - *"...researcher investigating gain wouldn't know if changes in scores would have occurred with no treatment anyway"*
- However, the hypothetical researcher in Lord (1967) is interested in gender differences and not in the effect of the diet
  - Lord (1967): "differential effect" of the diet
- under the Neyman-Rubin (a.k.a. potential outcomes) causal framework the effect of gender cannot be a causal research question

Neyman-Rubin framework

- Rubin, et al., (2003); Holland & Rubin (1983):
  - Suggest that Lord's example was a "poorly formulated causal assessment" since the potential outcome under the control diet is missing
  - *"...researcher investigating gain wouldn't know if changes in scores would have occurred with no treatment anyway"*
- However, the hypothetical researcher in Lord (1967) is interested in gender differences and not in the effect of the diet
  - Lord (1967): "differential effect" of the diet
- under the Neyman-Rubin (a.k.a. potential outcomes) causal framework the effect of gender cannot be a causal research question

Neyman-Rubin framework

- Rubin, et al., (2003); Holland & Rubin (1983):
    - Suggest that Lord’s example was a “poorly formulated causal assessment” since the potential outcome under the control diet is missing
    - *“...researcher investigating gain wouldn’t know if changes in scores would have occurred with no treatment anyway”*
- However, the hypothetical researcher in Lord (1967) is interested in gender differences and not in the effect of the diet
    - Lord (1967): “differential effect” of the diet
- under the Neyman-Rubin (a.k.a. potential outcomes) causal framework the effect of gender cannot be a causal research question

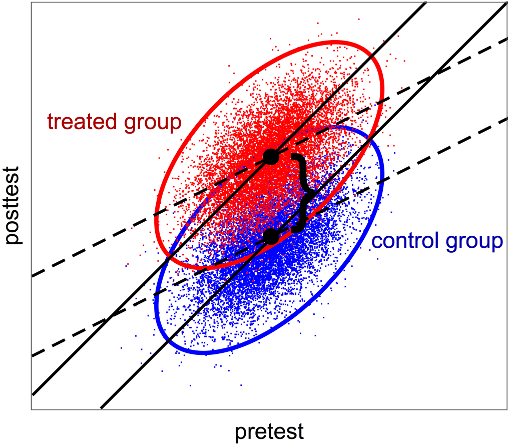
The two approaches

- CS approach:
  $Y_{post} - Y_{pre} = \beta_1 + \beta_2 G + \epsilon$
- RV approach:
  $Y_{post} = \beta_1 + \beta_2 G + \beta_3 Y_{pre} + \epsilon$

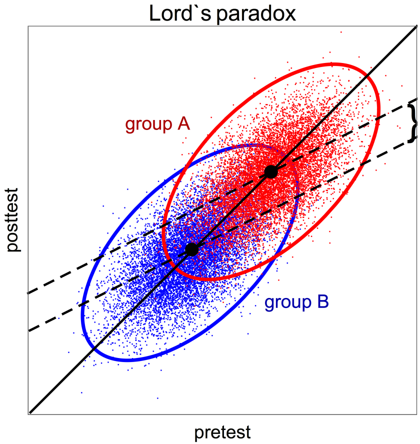
The two approaches

- CS approach:
  $Y_{post} - Y_{pre} = \beta_1 + \beta_2 G + \epsilon$
- RV approach:
  $Y_{post} = \beta_1 + \beta_2 G + \beta_3 Y_{pre} + \epsilon$

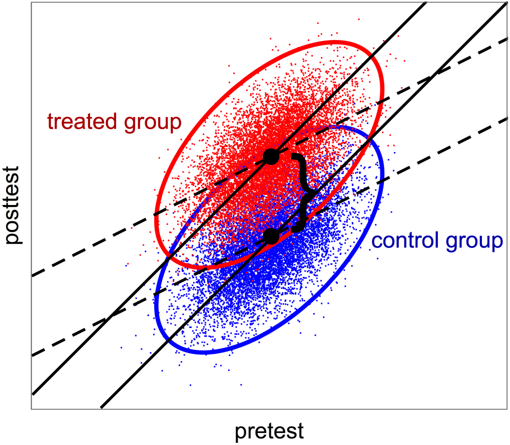
treatment assigned at random



treatment not at random



treatment assigned at random



Best of both?

- CS approach:
  $Y_{post} - Y_{pre} = \beta_1 + \beta_2 G + \epsilon$

- RV approach:
  $Y_{post} = \beta_1 + \beta_2 G + \beta_3 Y_{pre} + \epsilon$

Best of both?

- CS approach:
  $Y_{post} - Y_{pre} = \beta_1 + \beta_2 G + \epsilon$

- RV approach:
  $Y_{post} = \beta_1 + \beta_2 G + \beta_3 Y_{pre} + \epsilon$

Controlling for pretest in both?

- What if:
  $Y_{post} - Y_{pre} = \beta_1 + \beta_2 G + \beta_3^* Y_{pre} + \epsilon$
  $\downarrow$
  $Y_{post} = \beta_1 + \beta_2 G + (1 + \beta_3^*) Y_{pre} + \epsilon$
- RV approach:
  $Y_{post} = \beta_1 + \beta_2 G + \beta_3 Y_{pre} + \epsilon$

Controlling for pretest in both?

- What if:
  $Y_{post} - Y_{pre} = \beta_1 + \beta_2 G + \beta_3^* Y_{pre} + \epsilon$
  $\downarrow$
  $Y_{post} = \beta_1 + \beta_2 G + (1 + \beta_3^*) Y_{pre} + \epsilon$
- RV approach:
  $Y_{post} = \beta_1 + \beta_2 G + \beta_3 Y_{pre} + \epsilon$

RV approach

$$Y_2 = \beta_1 + \beta_2 G + \beta_3 Y_1 + \epsilon$$

- OLS yields unbiased estimates assuming:
  - ϵ uncorrelated with G and Y_1
  - correct specification
  - i.i.d.
  - no measurement error

RV approach

$$Y_2 = \beta_1 + \beta_2 G + \beta_3 Y_1 + \epsilon$$

- OLS yields unbiased estimates assuming:
  - ϵ uncorrelated with G and Y_1
  - correct specification
  - i.i.d.
  - no measurement error

RV approach

$$Y_2 = \beta_1 + \beta_2 G + \beta_3 Y_1 + \epsilon$$

- OLS yields unbiased estimates assuming:
  - ϵ uncorrelated with G and Y_1
  - correct specification
  - i.i.d.
  - no measurement error

RV approach

$$Y_2 = \beta_1 + \beta_2 G + \beta_3 Y_1 + \epsilon$$

- OLS yields unbiased estimates assuming:
  - ϵ uncorrelated with G and Y_1
  - correct specification
  - i.i.d.
  - no measurement error

RV approach

$$Y_2 = \beta_1 + \beta_2 G + \beta_3 Y_1 + \epsilon$$

- OLS yields unbiased estimates assuming:
  - ϵ uncorrelated with G and Y_1
  - correct specification
  - i.i.d.
  - no measurement error

RV approach

$$Y_2 = \beta_1 + \beta_2 G + \beta_3 Y_1 + \epsilon$$

- OLS yields unbiased estimates assuming:
  - ϵ uncorrelated with G and Y_1
  - correct specification
  - i.i.d.
  - no measurement error

CS approach

- assume H is binary: 1 if person ends up in the treatment group; 0 otherwise
- **pre:** $Y_1 = \beta_0 + \delta H + \epsilon_1$
  - δ: group differences that are stable
- **post:** $Y_2 = \beta_0 + \beta_1 + \delta H + \beta_2 G + \epsilon_2$
  - G is treatment indicator
  - $H = G$ (collinear)
  - β_1 represents the change that is occurring in both groups (e.g., gained knowledge during a school-year)
- $Y_2 - Y_1 = (\beta_0 - \beta_0) + \beta_1 + (\delta H - \delta H) + \beta_2 G + (\epsilon_2 - \epsilon_1)$
- $\Delta Y = \beta_1 + \beta_2 G + \epsilon^{\Delta}$
  - assuming ϵ^{Δ} is not correlated with G, OLS is consistent and hence the estimates are unbiased

CS approach

- assume H is binary: 1 if person ends up in the treatment group; 0 otherwise
- **pre:** $Y_1 = \beta_0 + \delta H + \epsilon_1$
  - δ: group differences that are stable
- **post:** $Y_2 = \beta_0 + \beta_1 + \delta H + \beta_2 G + \epsilon_2$
  - G is treatment indicator
  - $H = G$ (collinear)
  - β_1 represents the change that is occurring in both groups (e.g., gained knowledge during a school-year)
- $Y_2 - Y_1 = (\beta_0 - \beta_0) + \beta_1 + (\delta H - \delta H) + \beta_2 G + (\epsilon_2 - \epsilon_1)$
- $\Delta Y = \beta_1 + \beta_2 G + \epsilon^{\Delta}$
  - assuming ϵ^{Δ} is not correlated with G, OLS is consistent and hence the estimates are unbiased

CS approach

- assume H is binary: 1 if person ends up in the treatment group; 0 otherwise
- **pre:** $Y_1 = \beta_0 + \delta H + \epsilon_1$
    - δ: group differences that are stable
- **post:** $Y_2 = \beta_0 + \beta_1 + \delta H + \beta_2 G + \epsilon_2$
    - G is treatment indicator
    - $H = G$ (collinear)
    - β_1 represents the change that is occurring in both groups (e.g., gained knowledge during a school-year)
- $Y_2 - Y_1 = (\beta_0 - \beta_0) + \beta_1 + (\delta H - \delta H) + \beta_2 G + (\epsilon_2 - \epsilon_1)$
- $\Delta Y = \beta_1 + \beta_2 G + \epsilon^{\Delta}$
    - assuming ϵ^{Δ} is not correlated with G, OLS is consistent and hence the estimates are unbiased

CS approach

- assume H is binary: 1 if person ends up in the treatment group; 0 otherwise
- **pre:** $Y_1 = \beta_0 + \delta H + \epsilon_1$
  - δ: group differences that are stable
- **post:** $Y_2 = \beta_0 + \beta_1 + \delta H + \beta_2 G + \epsilon_2$
  - G is treatment indicator
  - $H = G$ (collinear)
  - β_1 represents the change that is occurring in both groups (e.g., gained knowledge during a school-year)
- $Y_2 - Y_1 = (\beta_0 - \beta_0) + \beta_1 + (\delta H - \delta H) + \beta_2 G + (\epsilon_2 - \epsilon_1)$
- $\Delta Y = \beta_1 + \beta_2 G + \epsilon^{\Delta}$
  - assuming ϵ^{Δ} is not correlated with G, OLS is consistent and hence the estimates are unbiased

CS approach

- assume H is binary: 1 if person ends up in the treatment group; 0 otherwise
- **pre:** $Y_1 = \beta_0 + \delta H + \epsilon_1$
  - δ: group differences that are stable
- **post:** $Y_2 = \beta_0 + \beta_1 + \delta H + \beta_2 G + \epsilon_2$
  - G is treatment indicator
  - $H = G$ (collinear)
  - β_1 represents the change that is occurring in both groups (e.g., gained knowledge during a school-year)
- $Y_2 - Y_1 = (\beta_0 - \beta_0) + \beta_1 + (\delta H - \delta H) + \beta_2 G + (\epsilon_2 - \epsilon_1)$
- $\Delta Y = \beta_1 + \beta_2 G + \epsilon^{\Delta}$
  - assuming ϵ^{Δ} is not correlated with G, OLS is consistent and hence the estimates are unbiased

CS approach

- assume H is binary: 1 if person ends up in the treatment group; 0 otherwise
- **pre:** $Y_1 = \beta_0 + \delta H + \epsilon_1$
    - δ: group differences that are stable
- **post:** $Y_2 = \beta_0 + \beta_1 + \delta H + \beta_2 G + \epsilon_2$
    - G is treatment indicator
    - $H = G$ (collinear)
    - β_1 represents the change that is occurring in both groups (e.g., gained knowledge during a school-year)
- $Y_2 - Y_1 = (\beta_0 - \beta_0) + \beta_1 + (\delta H - \delta H) + \beta_2 G + (\epsilon_2 - \epsilon_1)$
- $\Delta Y = \beta_1 + \beta_2 G + \epsilon^{\Delta}$
    - assuming ϵ^{Δ} is not correlated with G, OLS is consistent and hence the estimates are unbiased

CS approach

- assume H is binary: 1 if person ends up in the treatment group; 0 otherwise
- **pre:** $Y_1 = \beta_0 + \delta H + \epsilon_1$
  - δ: group differences that are stable
- **post:** $Y_2 = \beta_0 + \beta_1 + \delta H + \beta_2 G + \epsilon_2$
  - G is treatment indicator
  - $H = G$ (collinear)
  - β_1 represents the change that is occurring in both groups (e.g., gained knowledge during a school-year)
- $Y_2 - Y_1 = (\beta_0 - \beta_0) + \beta_1 + (\delta H - \delta H) + \beta_2 G + (\epsilon_2 - \epsilon_1)$
- $\Delta Y = \beta_1 + \beta_2 G + \epsilon^{\Delta}$
  - assuming ϵ^{Δ} is not correlated with G, OLS is consistent and hence the estimates are unbiased

CS approach

- assume H is binary: 1 if person ends up in the treatment group; 0 otherwise
- **pre:** $Y_1 = \beta_0 + \delta H + \epsilon_1$
  - δ: group differences that are stable
- **post:** $Y_2 = \beta_0 + \beta_1 + \delta H + \beta_2 G + \epsilon_2$
  - G is treatment indicator
  - $H = G$ (collinear)
  - β_1 represents the change that is occurring in both groups (e.g., gained knowledge during a school-year)
- $Y_2 - Y_1 = (\beta_0 - \beta_0) + \beta_1 + (\delta H - \delta H) + \beta_2 G + (\epsilon_2 - \epsilon_1)$
- $\Delta Y = \beta_1 + \beta_2 G + \epsilon^{\Delta}$
  - assuming ϵ^{Δ} is not correlated with G, OLS is consistent and hence the estimates are unbiased

CS approach

- assume H is binary: 1 if person ends up in the treatment group; 0 otherwise
- **pre:** $Y_1 = \beta_0 + \delta H + \epsilon_1$
  - δ: group differences that are stable
- **post:** $Y_2 = \beta_0 + \beta_1 + \delta H + \beta_2 G + \epsilon_2$
  - G is treatment indicator
  - $H = G$ (collinear)
  - β_1 represents the change that is occurring in both groups (e.g., gained knowledge during a school-year)
- $Y_2 - Y_1 = (\beta_0 - \beta_0) + \beta_1 + (\delta H - \delta H) + \beta_2 G + (\epsilon_2 - \epsilon_1)$
- $\Delta Y = \beta_1 + \beta_2 G + \epsilon^{\Delta}$
  - assuming ϵ^{Δ} is not correlated with G, OLS is consistent and hence the estimates are unbiased

CS approach

- assume H is binary: 1 if person ends up in the treatment group; 0 otherwise
- **pre:** $Y_1 = \beta_0 + \delta H + \epsilon_1$
  - δ: group differences that are stable
- **post:** $Y_2 = \beta_0 + \beta_1 + \delta H + \beta_2 G + \epsilon_2$
  - G is treatment indicator
  - $H = G$ (collinear)
  - β_1 represents the change that is occurring in both groups (e.g., gained knowledge during a school-year)
- $Y_2 - Y_1 = (\beta_0 - \beta_0) + \beta_1 + (\delta H - \delta H) + \beta_2 G + (\epsilon_2 - \epsilon_1)$
- $\Delta Y = \beta_1 + \beta_2 G + \epsilon^{\Delta}$
  - assuming ϵ^{Δ} is not correlated with G, OLS is consistent and hence the estimates are unbiased

Special case?

- Change score method:
  $Y_2 - Y_1 = \beta_1 + \beta_2 G + \epsilon^{\Delta}$
- rewrite:
  $Y_2 = \beta_1 + \beta_2 G + (1) Y_1 + \epsilon^{\Delta}$
- Many note that this is a special case of the:
  $Y_2 = \beta_1 + \beta_2 G + \beta_3 Y + \epsilon$
- see for instance:
  - Hedeker & Gibbons, 2006 (p. 8)
  - Van Breukelen, 2013, (p. 903)
  - Gelman & Hill, 2006 (p. 177)
    - *"an unnecessary assumption, namely that* $[\beta_3 = 1]$*"*

Special case?

- Change score method:
  $Y_2 - Y_1 = \beta_1 + \beta_2 G + \epsilon^{\Delta}$
- rewrite:
  $Y_2 = \beta_1 + \beta_2 G + (1) Y_1 + \epsilon^{\Delta}$
- Many note that this is a special case of the:
  $Y_2 = \beta_1 + \beta_2 G + \beta_3 Y + \epsilon$
- see for instance:
  - Hedeker & Gibbons, 2006 (p. 8)
  - Van Breukelen, 2013, (p. 903)
  - Gelman & Hill, 2006 (p. 177)
    - *"an unnecessary assumption, namely that [$\beta_3 = 1$]"*

Special case?

- Change score method:
  $Y_2 - Y_1 = \beta_1 + \beta_2 G + \epsilon^{\Delta}$
- rewrite:
  $Y_2 = \beta_1 + \beta_2 G + (1) Y_1 + \epsilon^{\Delta}$
- Many note that this is a special case of the:
  $Y_2 = \beta_1 + \beta_2 G + \beta_3 Y + \epsilon$
- see for instance:
  - Hedeker & Gibbons, 2006 (p. 8)
  - Van Breukelen, 2013, (p. 903)
  - Gelman & Hill, 2006 (p. 177)
    - *"an unnecessary assumption, namely that* $[\beta_3 = 1]$*"*

Special case?

- Change score method:
  $Y_2 - Y_1 = \beta_1 + \beta_2 G + \epsilon^{\Delta}$
- rewrite:
  $Y_2 = \beta_1 + \beta_2 G + (1) Y_1 + \epsilon^{\Delta}$
- Many note that this is a special case of the:

$$Y_2 = \beta_1 + \beta_2 G + \beta_3 Y + \epsilon$$

- see for instance:
  - Hedeker & Gibbons, 2006 (p. 8)
  - Van Breukelen, 2013, (p. 903)
  - Gelman & Hill, 2006 (p. 177)
    - *"an unnecessary assumption, namely, that $[\beta_3 = 1]$"*

Special case?

- Change score method:
  $Y_2 - Y_1 = \beta_1 + \beta_2 G + \epsilon^{\Delta}$
- rewrite:
  $Y_2 = \beta_1 + \beta_2 G + (1) Y_1 + \epsilon^{\Delta}$
- Many note that this is a special case of the:
  $Y_2 = \beta_1 + \beta_2 G + \beta_3 Y + \epsilon$
- see for instance:
  - Hedeker & Gibbons, 2006 (p. 8)
  - Van Breukelen, 2013, (p. 903)
  - Gelman & Hill, 2006 (p. 177)
    - *"an unnecessary assumption, namely, that* $[\beta_3 = 1]$*"*

Special case?

- Change score method:
  $Y_2 - Y_1 = \beta_1 + \beta_2 G + \epsilon^{\Delta}$
- rewrite:
  $Y_2 = \beta_1 + \beta_2 G + (1) Y_1 + \epsilon^{\Delta}$
- Many note that this is a special case of the:

  $Y_2 = \beta_1 + \beta_2 G + \beta_3 Y + \epsilon$
- see for instance:
  - Hedeker & Gibbons, 2006 (p. 8)
  - Van Breukelen, 2013, (p. 903)
  - Gelman & Hill, 2006 (p. 177)
    - *“an unnecessary assumption, namely, that* $[\beta_3 = 1]$*”*

Special case?

- Change score method:
  $Y_2 - Y_1 = \beta_1 + \beta_2 G + \epsilon^{\Delta}$
- rewrite:
  $Y_2 = \beta_1 + \beta_2 G + (1) Y_1 + \epsilon^{\Delta}$
- Many note that this is a special case of the:

  $Y_2 = \beta_1 + \beta_2 G + \beta_3 Y + \epsilon$
- see for instance:
  - Hedeker & Gibbons, 2006 (p. 8)
  - Van Breukelen, 2013, (p. 903)
  - Gelman & Hill, 2006 (p. 177)
    - *“an unnecessary assumption, namely, that* $[\beta_3 = 1]$*”*

Special case?

- Change score method:
  $Y_2 - Y_1 = \beta_1 + \beta_2 G + \epsilon^{\Delta}$
- rewrite:
  $Y_2 = \beta_1 + \beta_2 G + (1) Y_1 + \epsilon^{\Delta}$
- Many note that this is a special case of the:

  $Y_2 = \beta_1 + \beta_2 G + \beta_3 Y + \epsilon$
- see for instance:
  - Hedeker & Gibbons, 2006 (p. 8)
  - Van Breukelen, 2013, (p. 903)
  - Gelman & Hill, 2006 (p. 177)
    - *"an unnecessary assumption, namely, that* $[\beta_3 = 1]$*"*

Special case?

- Change score method:
  $Y_2 - Y_1 = \beta_1 + \beta_2 G + \epsilon^{\Delta}$
- rewrite:
  $Y_2 = \beta_1 + \beta_2 G + (1) Y_1 + \epsilon^{\Delta}$
- Many note that this is a special case of the:
  $Y_2 = \beta_1 + \beta_2 G + \beta_3 Y + \epsilon$
- see for instance:
  - Hedeker & Gibbons, 2006 (p. 8)
  - Van Breukelen, 2013, (p. 903)
  - Gelman & Hill, 2006 (p. 177)
    - *"an unnecessary assumption, namely, that* $[\beta_3 = 1]$"

Special case?

- $Y_2 = \beta_1 + \beta_2 G + (1) Y_1 + \epsilon^{\Delta}$
- $Y_2 = \beta_1 + \beta_2 G + \beta_3 Y_1 + \epsilon^{\Delta}$
  - inconsistent estimates since ϵ^{Δ} is negatively correlated with Y_1 by construction
    - $Y_2 - Y_1 = \beta_1 + \beta_2 G + \epsilon^{\Delta}$
- not a special case
  - The two approaches represent two completely different models!
  - Overlooking this crucial distinction has been the most common error in comparisons of the two approaches
  - Any discussion of Lord's paradox that does not acknowledge this distinction is likely to be misleading

Special case?

- $Y_2 = \beta_1 + \beta_2 G + (1) Y_1 + \epsilon^{\Delta}$
- $Y_2 = \beta_1 + \beta_2 G + \beta_3 Y_1 + \epsilon^{\Delta}$
  - inconsistent estimates since ϵ^{Δ} is negatively correlated with Y_1 by construction
    - $Y_2 - Y_1 = \beta_1 + \beta_2 G + \epsilon^{\Delta}$
- not a special case
  - The two approaches represent two completely different models!
  - Overlooking this crucial distinction has been the most common error in comparisons of the two approaches
  - Any discussion of Lord's paradox that does not acknowledge this distinction is likely to be misleading

Special case?

- $Y_2 = \beta_1 + \beta_2 G + (1) Y_1 + \epsilon^{\Delta}$
- $Y_2 = \beta_1 + \beta_2 G + \beta_3 Y_1 + \epsilon^{\Delta}$
    - inconsistent estimates since ϵ^{Δ} is negatively correlated with Y_1 by construction
        - $Y_2 - Y_1 = \beta_1 + \beta_2 G + \epsilon^{\Delta}$
- not a special case
    - The two approaches represent two completely different models!
    - Overlooking this crucial distinction has been the most common error in comparisons of the two approaches
    - Any discussion of Lord's paradox that does not acknowledge this distinction is likely to be misleading

Special case?

- $Y_2 = \beta_1 + \beta_2 G + (1) Y_1 + \epsilon^{\Delta}$
- $Y_2 = \beta_1 + \beta_2 G + \beta_3 Y_1 + \epsilon^{\Delta}$
  - inconsistent estimates since ϵ^{Δ} is negatively correlated with Y_1 by construction
    - $Y_2 - Y_1 = \beta_1 + \beta_2 G + \epsilon^{\Delta}$
- not a special case
  - The two approaches represent two completely different models!
  - Overlooking this crucial distinction has been the most common error in comparisons of the two approaches
  - Any discussion of Lord's paradox that does not acknowledge this distinction is likely to be misleading

Special case?

- $Y_2 = \beta_1 + \beta_2 G + (1) Y_1 + \epsilon^{\Delta}$
- $Y_2 = \beta_1 + \beta_2 G + \beta_3 Y_1 + \epsilon^{\Delta}$
  - inconsistent estimates since ϵ^{Δ} is negatively correlated with Y_1 by construction
    - $Y_2 - Y_1 = \beta_1 + \beta_2 G + \epsilon^{\Delta}$
- not a special case
  - The two approaches represent two completely different models!
  - Overlooking this crucial distinction has been the most common error in comparisons of the two approaches
  - Any discussion of Lord's paradox that does not acknowledge this distinction is likely to be misleading

Special case?

- $Y_2 = \beta_1 + \beta_2 G + (1) Y_1 + \epsilon^{\Delta}$
- $Y_2 = \beta_1 + \beta_2 G + \beta_3 Y_1 + \epsilon^{\Delta}$
  - inconsistent estimates since ϵ^{Δ} is negatively correlated with Y_1 by construction
    - $Y_2 - Y_1 = \beta_1 + \beta_2 G + \epsilon^{\Delta}$
- not a special case
  - The two approaches represent two completely different models!
  - Overlooking this crucial distinction has been the most common error in comparisons of the two approaches
  - Any discussion of Lord's paradox that does not acknowledge this distinction is likely to be misleading

Special case?

- $Y_2 = \beta_1 + \beta_2 G + (1) Y_1 + \epsilon^{\Delta}$
- $Y_2 = \beta_1 + \beta_2 G + \beta_3 Y_1 + \epsilon^{\Delta}$
  - inconsistent estimates since ϵ^{Δ} is negatively correlated with Y_1 by construction
    - $Y_2 - Y_1 = \beta_1 + \beta_2 G + \epsilon^{\Delta}$
- not a special case
  - The two approaches represent two completely different models!
  - Overlooking this crucial distinction has been the most common error in comparisons of the two approaches
  - Any discussion of Lord's paradox that does not acknowledge this distinction is likely to be misleading

Special case?

- $Y_2 = \beta_1 + \beta_2 G + (1) Y_1 + \epsilon^{\Delta}$
- $Y_2 = \beta_1 + \beta_2 G + \beta_3 Y_1 + \epsilon^{\Delta}$
  - inconsistent estimates since ϵ^{Δ} is negatively correlated with Y_1 by construction
    - $Y_2 - Y_1 = \beta_1 + \beta_2 G + \epsilon^{\Delta}$
- not a special case
  - The two approaches represent two completely different models!
  - Overlooking this crucial distinction has been the most common error in comparisons of the two approaches
  - Any discussion of Lord's paradox that does not acknowledge this distinction is likely to be misleading

Which approach should I use?

- It depends...
- When randomized: both are fine, CS has less power
- When the treatment is assigned based on the pretest:
  – it becomes necessary to control for the pretest

Which approach should I use?

- It depends...
- When randomized: both are fine, CS has less power
- When the treatment is assigned based on the pretest:
  – it becomes necessary to control for the pretest

Which approach should I use?

- It depends...
- When randomized: both are fine, CS has less power
- When the treatment is assigned based on the pretest:
  – it becomes necessary to control for the pretest

Which approach should I use?

- Regression to the mean
  - RV and CS approaches assume regressions to "different means"
  - The RV approach assumes that two groups will regress toward the grand mean
  - The CS approach assumes that the posttest scores will regress to their group-specific means

Which approach should I use?

- Regression to the mean
  - RV and CS approaches assume regressions to "different means"
  - The RV approach assumes that two groups will regress toward the grand mean
  - The CS approach assumes that the posttest scores will regress to their group-specific means

Which approach should I use?

- Regression to the mean
    - RV and CS approaches assume regressions to "different means"
    - The RV approach assumes that two groups will regress toward the grand mean
    - The CS approach assumes that the posttest scores will regress to their group-specific means

Which approach should I use?

- Regression to the mean
  - RV and CS approaches assume regressions to "different means"
  - The RV approach assumes that two groups will regress toward the grand mean
  - The CS approach assumes that the posttest scores will regress to their group-specific means

Two different conclusions

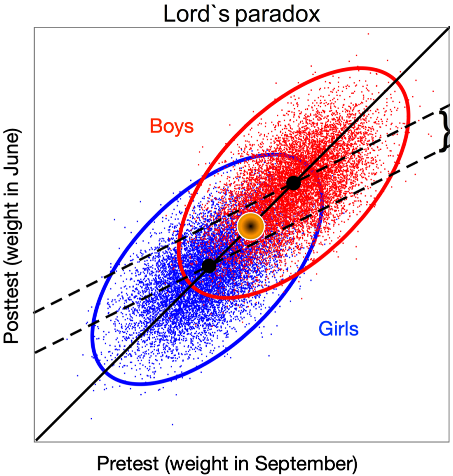
Statistician 1

- For girls: no difference between weights in September and June
- For boys: no difference between weights in September and June
- Conclusion: no differences between boys and girls in weight gain.

Statistician 2

- "... boys showed significantly more gain in weight than the girls when proper allowance is made for differences in initial weight between the two sexes"
- Adjusting/controlling for weight in September, boys are higher in their weight in June.
- Conclusion: boys showed more gain than girls.

Lord's paradox



Which approach should I use?

- When is the assumption of "regression toward the grand mean" plausible?
- The CS approach becomes a very robust choice in cases when this assumption is too heroic
- However, the CS approach can give misleading results too: if two groups are indeed regressing to the grand mean
  - The group with the lower mean will tend to gain more than the group with the higher mean due to the "regression to the mean" reality
  - The difference in gains between groups will simply be the result of the "regression artifact"
- Grand mean vs. group-mean: how do I know?
- Are groups exchangeable?

Which approach should I use?

- When is the assumption of "regression toward the grand mean" plausible?
- The CS approach becomes a very robust choice in cases when this assumption is too heroic
- However, the CS approach can give misleading results too: if two groups are indeed regressing to the grand mean
    - The group with the lower mean will tend to gain more than the group with the higher mean due to the "regression to the mean" reality
    - The difference in gains between groups will simply be the result of the "regression artifact"
- Grand mean vs. group-mean: how do I know?
- Are groups exchangeable?

Which approach should I use?

- When is the assumption of "regression toward the grand mean" plausible?
- The CS approach becomes a very robust choice in cases when this assumption is too heroic
- However, the CS approach can give misleading results too: if two groups are indeed regressing to the grand mean
    - The group with the lower mean will tend to gain more than the group with the higher mean due to the "regression to the mean" reality
    - The difference in gains between groups will simply be the result of the "regression artifact"
- Grand mean vs. group-mean: how do I know?
- Are groups exchangeable?

Which approach should I use?

- When is the assumption of "regression toward the grand mean" plausible?
- The CS approach becomes a very robust choice in cases when this assumption is too heroic
- However, the CS approach can give misleading results too: if two groups are indeed regressing to the grand mean
  - The group with the lower mean will tend to gain more than the group with the higher mean due to the "regression to the mean" reality
  - The difference in gains between groups will simply be the result of the "regression artifact"
- Grand mean vs. group-mean: how do I know?
- Are groups exchangeable?

Which approach should I use?

- When is the assumption of "regression toward the grand mean" plausible?
- The CS approach becomes a very robust choice in cases when this assumption is too heroic
- However, the CS approach can give misleading results too: if two groups are indeed regressing to the grand mean
  - The group with the lower mean will tend to gain more than the group with the higher mean due to the "regression to the mean" reality
  - The difference in gains between groups will simply be the result of the "regression artifact"
- Grand mean vs. group-mean: how do I know?
- Are groups exchangeable?

Which approach should I use?

- When is the assumption of “regression toward the grand mean” plausible?
- The CS approach becomes a very robust choice in cases when this assumption is too heroic
- However, the CS approach can give misleading results too: if two groups are indeed regressing to the grand mean
  - The group with the lower mean will tend to gain more than the group with the higher mean due to the “regression to the mean” reality
  - The difference in gains between groups will simply be the result of the “regression artifact”
- Grand mean vs. group-mean: how do I know?
- Are groups exchangeable?

Which approach should I use?

- When is the assumption of "regression toward the grand mean" plausible?
- The CS approach becomes a very robust choice in cases when this assumption is too heroic
- However, the CS approach can give misleading results too: if two groups are indeed regressing to the grand mean
  - The group with the lower mean will tend to gain more than the group with the higher mean due to the "regression to the mean" reality
  - The difference in gains between groups will simply be the result of the "regression artifact"
- Grand mean vs. group-mean: how do I know?
- Are groups exchangeable?

Exchangeability assumption

- Inference must be based on careful specification of the relevant subpopulation
- Are groups exchangeable with respect to the outcome of interest?
  - It might be that Group A and Group B are exchangeable if the outcome of interest is the mean number of hours spent in the gym
  - And not exchangeable if the outcome of interest is the mean number of calories burned
  - Subject matter expertise is necessary to answer this question
- Exchangeability is guaranteed by randomization

Exchangeability assumption

- Inference must be based on careful specification of the relevant subpopulation
- Are groups exchangeable with respect to the outcome of interest?
    - It might be that Group A and Group B are exchangeable if the outcome of interest is the mean number of hours spent in the gym
    - And not exchangeable if the outcome of interest is the mean number of calories burned
    - Subject matter expertise is necessary to answer this question
- Exchangeability is guaranteed by randomization

Exchangeability assumption

- Inference must be based on careful specification of the relevant subpopulation
- Are groups exchangeable with respect to the outcome of interest?
  - It might be that Group A and Group B are exchangeable if the outcome of interest is the mean number of hours spent in the gym
  - And not exchangeable if the outcome of interest is the mean number of calories burned
  - Subject matter expertise is necessary to answer this question
- Exchangeability is guaranteed by randomization

Exchangeability assumption

- Inference must be based on careful specification of the relevant subpopulation
- Are groups exchangeable with respect to the outcome of interest?
    - It might be that Group A and Group B are exchangeable if the outcome of interest is the mean number of hours spent in the gym
    - And not exchangeable if the outcome of interest is the mean number of calories burned
    - Subject matter expertise is necessary to answer this question
- Exchangeability is guaranteed by randomization

Exchangeability assumption

- Inference must be based on careful specification of the relevant subpopulation
- Are groups exchangeable with respect to the outcome of interest?
    - It might be that Group A and Group B are exchangeable if the outcome of interest is the mean number of hours spent in the gym
    - And not exchangeable if the outcome of interest is the mean number of calories burned
    - Subject matter expertise is necessary to answer this question
- Exchangeability is guaranteed by randomization

Exchangeability assumption

- Inference must be based on careful specification of the relevant subpopulation
- Are groups exchangeable with respect to the outcome of interest?
    - It might be that Group A and Group B are exchangeable if the outcome of interest is the mean number of hours spent in the gym
    - And not exchangeable if the outcome of interest is the mean number of calories burned
    - Subject matter expertise is necessary to answer this question
- Exchangeability is guaranteed by randomization

Measurement error and scale

- So far: pretest and the posttest measures do not contain any measurement error

- What if there is a measurement error in pre and post?
  - RV approach will produce biased estimates

- What if pre and post have different scales?
  - CS approach requires a common scale

- What if my data is clustered?
  - CS approach simplifies things

Measurement error and scale

- So far: pretest and the posttest measures do not contain any measurement error
- What if there is a measurement error in pre and post?
  - RV approach will produce biased estimates
- What if pre and post have different scales?
  - CS approach requires a common scale
- What if my data is clustered?
  - CS approach simplifies things

Measurement error and scale

- So far: pretest and the posttest measures do not contain any measurement error
- What if there is a measurement error in pre and post?
  - RV approach will produce biased estimates
- What if pre and post have different scales?
  - CS approach requires a common scale
- What if my data is clustered?
  - CS approach simplifies things

Measurement error and scale

- So far: pretest and the posttest measures do not contain any measurement error
- What if there is a measurement error in pre and post?
  - RV approach will produce biased estimates
- What if pre and post have different scales?
  - CS approach requires a common scale
- What if my data is clustered?
  - CS approach simplifies things

Measurement error and scale

- So far: pretest and the posttest measures do not contain any measurement error
- What if there is a measurement error in pre and post?
    - RV approach will produce biased estimates
- What if pre and post have different scales?
    - CS approach requires a common scale
- What if my data is clustered?
    - CS approach simplifies things

Measurement error and scale

- So far: pretest and the posttest measures do not contain any measurement error

- What if there is a measurement error in pre and post?
  - RV approach will produce biased estimates

- What if pre and post have different scales?
  - CS approach requires a common scale

- What if my data is clustered?
  - CS approach simplifies things

Measurement error and scale

- So far: pretest and the posttest measures do not contain any measurement error
- What if there is a measurement error in pre and post?
  - RV approach will produce biased estimates
- What if pre and post have different scales?
  - CS approach requires a common scale
- What if my data is clustered?
  - CS approach simplifies things

What is next?

- Typical question: does the treatment have an effect?
- Tools: qualitative insights + quantitative analysis + healthy dose of skepticism
- None of the statistical methods, no matter how fancy and sophisticated they are, will be able to compensate for the sloppy study design (David Freedman)

What is next?

- Typical question: does the treatment have an effect?
- Tools: qualitative insights + quantitative analysis + healthy dose of skepticism
- None of the statistical methods, no matter how fancy and sophisticated they are, will be able to compensate for the sloppy study design (David Freedman)

What is next?

- Typical question: does the treatment have an effect?
- Tools: qualitative insights + quantitative analysis + healthy dose of skepticism
- None of the statistical methods, no matter how fancy and sophisticated they are, will be able to compensate for the sloppy study design (David Freedman)

thank you

questions?

perman@berkeley.edu
markw@berkeley.edu