Randomization Inference with An Instrumental Variable: Two Examples and Some Theory

Paul R. Rosenbaum, Department of Statistics, Wharton School
University of Pennsylvania, Philadelphia, PA 19104-6340 US

Talk based on several papers, including one with Guido Imbens and another with Robert Greevy, Jeffrey Silber, and Avital Cnaan.

Abstract

An instrument partially manipulates a treatment, but the instrument affects the response only indirectly through its manipulation of the treatment. Two examples of randomization inference with an instrumental variable will be discussed: a randomized trial with a strong instrument, and a widely discussed study from labor economics with a clever but extremely weak instrument. The randomized, placebo controlled trial concerns an attempt to use a drug, enalapril, to preserve cardiac function of children with cancer who had been treated with anthracyclines. As is often true in clinical trials, compliance with the protocol was imperfect: some children consumed less than the prescribed dose of drug. In this case, the randomization is the instrument, manipulating but not fully controlling the dose of drug received. The example from labor economics is due to Angrist and Krueger, who wished to study the economic returns to additional years of education. In the US, laws concerning compulsory education require children to remain in school until a particular birthday, but the school year begins in September for all children, so these laws require different students to attend school for slightly different periods of time based on their date of birth. In their study, Angrist and Krueger used quarter-of-birth as an instrument for the treatment, years of schooling. The instrument is, of course, extremely weak, because it adds at most a fraction of a year to schooling for a fairly small fraction of all students.

In the randomized trial, randomization inference with an instrumental variable permits randomization to form the 'reasoned basis for inference,' in Fisher's phrase, entirely avoiding biases from self-selection in the decision to comply with the study protocol. Moreover, this analysis agrees with the so-called 'intent-to-treat' analysis, in the sense that both analyses report exactly the same significance level for testing the null hypothesis of no treatment effect.

A large literature shows that conventional methods give incorrect conclusions with weak instruments; for instance, 95% confidence intervals do not cover 95% of the time. In contrast, randomization inference does not have this problem, even in situations that are not identified: its 95% confidence intervals do cover 95% of the time, with intervals whose (possibly infinite) length appropriately reflect the strength of the instrument. Moreover, theory shows this is the only way to achieve such coverage in the absence of distributional assumptions. This is illustrated using the study from labor economics and simulations.

Randomization Inference with An Instrumental Variable:

Two Examples and Some Theory

Paul R. Rosenbaum, University of Pennsylvania

- Greevy, R., Silber, J., Cnaan, A., Rosenbaum, P. R. (2004) Randomization inference with imperfect compliance in the ACE-inhibitor after anthracycline randomized trial. *Journal of the American Statistical Association*, **99**: 7-15.

- Imbens, G. and Rosenbaum, P. R. (2005) Robust, accurate confidence intervals with a weak instrument: Quarter of birth and education. *Journal of the Royal Statistical Society*, A, 168, Part 1, to appear.

(All analyses are intended to illustrate statistical methods, not to reach subject matter conclusions.)

1 Three IV Problems

- *Noncompliance in randomized trials*. Randomization inference with IV permits strict adherence to the logic of randomized controlled trials, while permitting allowance for (typically nonrandom) noncompliance.

- *Weak instruments*: Conventional two-stage least squares works poorly: 95% confidence intervals cover much less than 95% of the time. It's really two problems, not one:

1. Weak instruments may (or may not) provide limited information.

2. Two-stage least squares often exaggerates the information provided.

Randomization inference with IV fixes problem 2, thereby clarifying problem 1.

2 Outline of Talk

1. Randomized trial with noncompliance. Comparing intent-to-treat with randomization inference with an instrumental variable.

2. A study from labor economics with a weak instrument.

3 The AAA Randomized Trial

- About $\frac{2}{3}$ of childhood cancers are cured, often using a class of drugs called anthracyclines.

- Although fairly effective against cancer, these drugs may damage the heart. Can damage be limited by another drug, enalapril?

- ACE-Inhibitor After Anthracycline (AAA) Randomized Trial (Silber, et al. 2001) concerned children under age 20:

1. who had survived at least 4 years after cancer diagnosis and at least 2 years after the completion of all cancer treatment,

2. and who had certain defined forms of decline in cardiac systolic performance after treatment with an anthracycline.

4 Structure of the Trial

- 135 children were randomly assigned to enalapril or placebo.

- Cardiac performance measured prior to treatment and at six months intervals thereafter.

- Began in October 1994, closed enrollment in March 1999, and collected data until March 2001, so the final patient to be enrolled could be followed for two years, and many patients were observed for much longer.

5 Randomization and Noncompliance

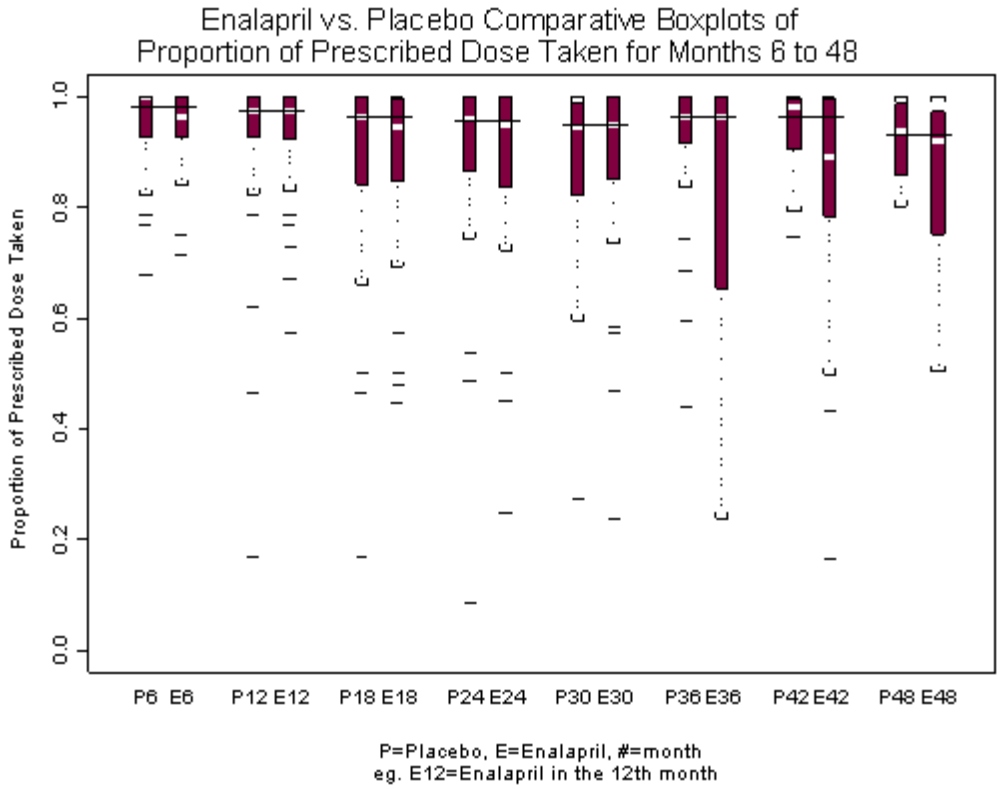
- In his *Design of Experiments*, Sir Ronald Fisher (1935) showed that random assignment of treatment justified or formed the 'reasoned basis for inference' about the effects caused by the treatment.

- That argument describes the situation in which patients assigned to a treatment actually receive it.

- In AAA and many trials, compliance with protocol is imperfect. Many children took less than the assigned dose of drug.

by time

pw50



1.*pdf*

Compliance with enalapril and placebo do not differ significantly as judged by Wei & Lachin (1984) test.

6 Options for Analysis

- Ignore noncompliance: The intent-to-treat analysis. Fully justified by randomization, but concerns the effect of assigning a patient to take a drug rather than the effect of taking it.

- Ignore assignment: Various analyses focus on the doses actually received. These are not justified by randomization, may face self-selection biases similar to those found in nonrandomized observational studies.

- Use both assignment and dose received in different roles in an instrumental variable analysis.

7 What Can Go Wrong?

- Example from May, et al. (1981): Coronary Drug Project. A randomized trial of lipid lowering drugs for individuals who had survived a myocardial infarction.

- Good compliers in the group assigned to clofibrate had a five year mortality rate of 15.0%, whereas poor compliers had a mortality rate of 24.6%, so clofibrate looks promising.

- In the placebo group, good compliers had a mortality rate of 15.1% and poor compliers had a mortality rate of 28.2%, so placebo looks better than clofibrate.

- A comparison of groups defined in terms of compliance is a comparison of self-selected groups, which may differ in outcomes for reasons that are not the effects of the treatment.

8 Notation for Treatment Assignment

- S strata defined by pretreatment covariates, $s = 1, \ldots, S$, with n_s subjects in stratum s, and $N = \sum n_s$ subjects in total

- $Z_{si} = 1$ if the i^{th} subject in stratum s is assigned to treatment, $Z_{si} = 0$ if assigned to control. $\mathbf{Z} = \left(Z_{11}, Z_{12}, \ldots, Z_{S,n_S}\right)^T$

- $m_s = \sum_{i=1}^{n_s} Z_{si}$ treated subjects in stratum s.

- Eg, treated/control matched pairs is the special case with $n_s = 2$, $m_s = 1$ for $s = 1, \ldots, S$

- In AAA, $S = 1$, $N = n_1 = 135$, $m_1 = 69$.

9 Random Assignment

- There are $K = \prod_{s=1}^{S} \binom{n_s}{m_s}$ possible values $\mathbf{z}$ of the $N-$dimensional treatment assignment $\mathbf{Z}$ with $m_s = \sum_{i=1}^{n_s} z_{si}$ for $s = 1, \ldots, S$.

- Collect these K vectors $\mathbf{z}$ in the set Ω.

- In a randomized experiment, $\mathbf{Z}$ is picked at random from Ω, that is, $\mathsf{Pr}(\mathbf{Z} = \mathbf{z}) = \frac{1}{K}$ for each $\mathbf{z} \in \Omega$.

- Randomization inference uses $\mathsf{Pr}(\mathbf{Z} = \mathbf{z}) = \frac{1}{K}$ as the basis for inference. Quantities that depend on $\mathbf{Z}$ are random variables; other quantities are fixed features of the finite population of N subjects.

10 Responses

- People are observed for *up to* K time periods, $k = 1, \ldots, K$, with $K = 13$ in AAA.

- Responses for later periods k are very often missing because of the date of entry into the study.

- Under treatment, $Z_{si} = 1$, person i in stratum s would have responses y_{Tsi1}, ..., y_{TsiK}, and under control, $Z_{si} = 0$, responses y_{Csi1}, ..., y_{CsiK}; eg Neyman (1923), Rubin (1974). (Fixed)

- Censoring due to late entry implies that y_{Tsik} would be missing under treatment if and only if y_{Csik} would be missing under control.

- Observed response from this person is Y_{si1}, ..., Y_{siK} where $Y_{sik} = Z_{si}\, y_{Tsik} + (1 - Z_{si})\, y_{Csik}$ if this person is observed for k periods and is missing otherwise. (Random)

11 Test of No Treatment Effect

- Null hypothesis of no treatment effect is: H_0 : $y_{Tsi1} = y_{Csi1}$, ..., $y_{TsiK} = y_{CsiK}$ for all s, i.

- If Y_{sjk} is missing due to the date of entry, H_0 is true for patient s, i at this k.

- Wei and Lachin (1984) proposed a test based on

$$\begin{aligned} U_{sijk} &= 1 \; if \; Y_{sik} > Y_{sjk} \\ &= -1 \; if \; Y_{sik} < Y_{sjk} \\ &= 0 \; if \; Y_{sik} = Y_{sjk} \; or \; if \; either \; is \; missing, \end{aligned} \tag{1}$$

so that $T_{sk} = \sum_{i=1}^{I} \sum_{j=1}^{I} Z_{si} \left(1 - Z_{sj}\right) U_{sijk}$. U_{sijk} is fixed under H_0.

- The Wei and Lachin statistic is $T = \sum_s \sum_k T_{sk}$.

12 T is a Linear Rank Statistic

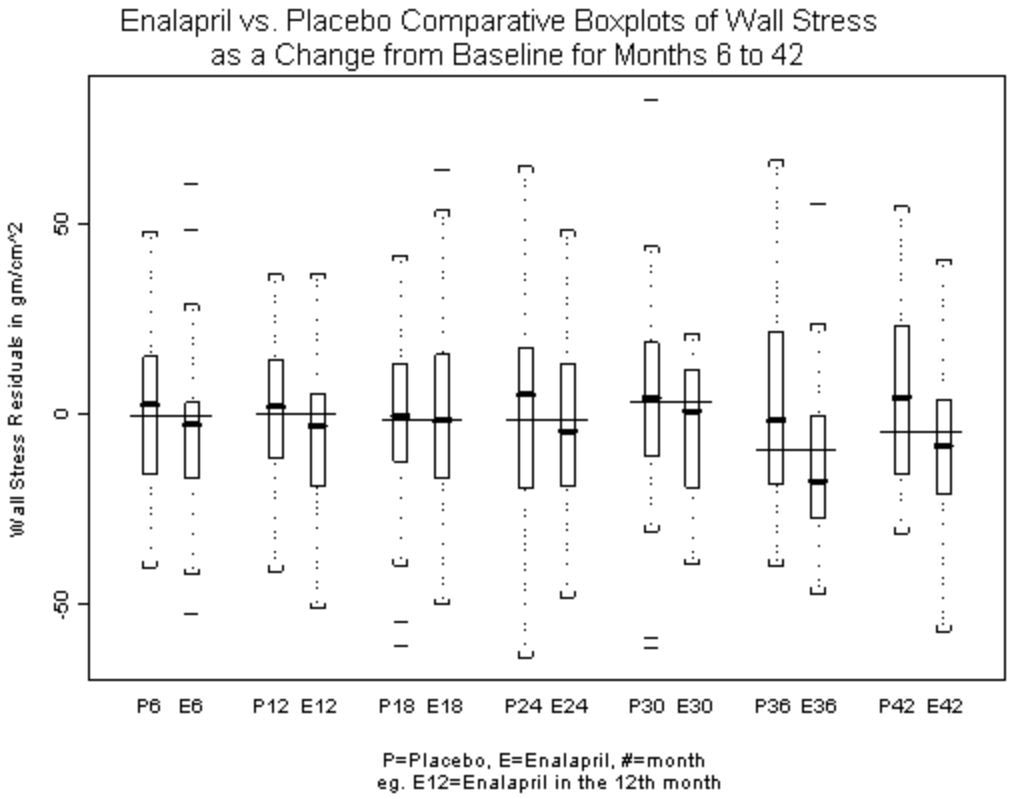
Because $Z_{si}Z_{sj}U_{sijk} = -Z_{sj}Z_{si}U_{sjik}$, it follows that $\sum_{i=1}^{I}\sum_{j=1}^{I} Z_{si}Z_{sj}U_{sijk} = 0$ so that

$$\begin{aligned} T &= \sum_{i,j,k,s} Z_{si}\left(1 - Z_{sj}\right)U_{sijk} && (2) \\ &= \sum_{i,s} Z_{si}\, q_{si} && (3) \end{aligned}$$

where

$$q_{si} = \sum_{j,k} U_{ijk}. \qquad (4)$$

so under H_0, the statistic T is the sample total in a stratified random sample from a finite population.



Horizontal lines are the pooled median at each time k.

Intent-to-treat analysis is essentially an analysis of these data, ignoring the pill counts.

13 Intent-to-Treat Analysis

- Compares children assigned to enalapril and placebo, ignoring doses, estimates typical effect.

- Model of constant effect: $y_{Tsi1} = y_{Csi1} + \tau$, ..., $y_{TsiK} = y_{CsiK} + \tau$ for all s, i.

- Test $H_0 : \tau = \tau_0$ by calculating $A_{sik} = Y_{sik} - Z_{si}\,\tau_0$ which equals y_{Csik} under H_0, and testing the null hypothesis of no effect for A_{sik}.

- For $K = 1$, yields usual Hodges-Lehmann estimate of shift for Wilcoxon's rank sum.

Intent-to-Treat			
P-value for no effect		0.0449	
Hodges-Lehmann Estimate		−6.57	
95% Confidence Interval	−12.60		−0.10

(Δ Wall Stress $\frac{gm}{cm^2}$)

14 Doses

- With noncompliance, the doses actually received are outcomes of the randomly assigned encouragement to take enalapril or placebo.

- Analogously, under treatment, $Z_{si} = 1$, person i in stratum s would have doses of enalapril d_{Tsi1}, ..., d_{TsiK}, and under control, $Z_{si} = 0$, doses d_{Csi1}, ..., d_{CsiK}.

- In AAA, $d_{Csi1} = \ldots = d_{CsiK} = 0$, but this is not essential to the argument.

- Observed dose from this person is D_{si1}, ..., D_{siK} where $D_{sik} = Z_{si}\, d_{Tsik} + (1 - Z_{si})\, d_{Csik}$ if this person is observed for k periods and is missing otherwise.

15 Effect Proportional to Dose

- Model says effect caused by enalapril is proportional to dose consumed:

$$y_{Tsik} - y_{Csik} = \beta \left(d_{Tsik} - d_{Csik} \right) \qquad (5)$$

as distinct from the constant effect model used previously with

$$y_{Tsik} - y_{Csik} = \tau.$$

- In AAA, $d_{Csik} = 0$ enalapril consumed in the placebo group. Full compliance in the enalapril group is $d_{Tsik} = 1$.

- A person who refuses all medication, $d_{Csik} = d_{Tsik} = 0$, has effect

$$y_{Tsik} - y_{Csik} = \beta \left(0 - 0 \right) = 0$$

whereas a person who is fully compliant with the protocol, $d_{Tsik} = 1$, has effect

$$y_{Tsik} - y_{Csik} = \beta \left(1 - 0 \right) = \beta$$

16 Exclusion Restriction

- The model

$$y_{Tsik} - y_{Csik} = \beta \left(d_{Tsik} - d_{Csik} \right)$$

  embodies the exclusion restriction.

- Being assigned to enalapril rather than placebo affects the response, but only indirectly by affecting shifting the amount of enalapril consumed from d_{Csik} to d_{Tsik}.

- The random numbers that assign people to enalapril rather than placebo have no effect on wall stress except to the extent that they influence the amount of enalapril consumed.

- A critical assumption of IV, albeit a plausible one here.

17 Randomization Test with IV

- If effect is proportional to dose,

$$y_{Tsik} - y_{Csik} = \beta\,(d_{Tsik} - d_{Csik})$$

then

$$\begin{aligned} y_{Tsik} - \beta\, d_{Tsik} &= y_{Csik} - \beta\, d_{Csik} \\ &= a_{sik}, \text{ say} \end{aligned}$$

That is, the a_{sik} are not affected by the treatment. (Actually, in AAA, $d_{Csik} = 0$, so $y_{Tsik} - \beta\, d_{Tsik} = y_{Csik} = a_{sik}$.)

- If $H_0 : \beta = \beta_0$ were true, then

$$\begin{aligned} Y_{sik} - \beta_0\, D_{sik} &= y_{Tsik} - \beta\, d_{Tsik} = a_{sik} \text{ if } Z_{si} = 1 \\ &= y_{Csik} - \beta\, d_{Csik} = a_{sik} \text{ if } Z_{si} = 0 \end{aligned}$$

- Test $H_0 : \beta = \beta_0$ by applying Wei/Lachin test to $Y_{sik} - \beta_0\, D_{sik}$. Invert for HL estimates and confidence intervals.

18 Remarks

- Intent-to-treat and IV compare the same test statistic to the same reference distribution created by randomization.

- They differ in that intent-to-treat applies the test to $Y_{sik} - \tau_0 Z_{si}$ to test

$$y_{Tsik} - y_{Csik} = \tau_0$$

while IV applies the test to $Y_{sik} - \beta_0 D_{sik}$ to test

$$y_{Tsik} - y_{Csik} = \beta_0 \left(d_{Tsik} - d_{Csik}\right).$$

Same logic, different null hypotheses.

- For no effect, $H_0 : y_{Tsik} - y_{Csik} = 0$, the two tests are identical. They agree exactly about whether no effect is plausible.

- Both tests use (i) random assignment of treatments, and (ii) the null hypothesis being tested.

19 Results Compared

Intent-to-Treat			
P-value for no effect		0.0449	
Hodges-Lehmann Estimate		-6.57	
95% Confidence Interval	-12.60		-0.10
90% Confidence Interval	-11.70		-1.09
$\frac{2}{3}$ Confidence Interval	-8.00		-5.07
Instrumental Variable			
P-value for no effect		0.0449	
Hodges-Lehmann Estimate		-7.11	
95% Confidence Interval	-13.82		-0.11
90% Confidence Interval	-12.70		-1.14
$\frac{2}{3}$ Confidence Interval	-8.81		-5.65

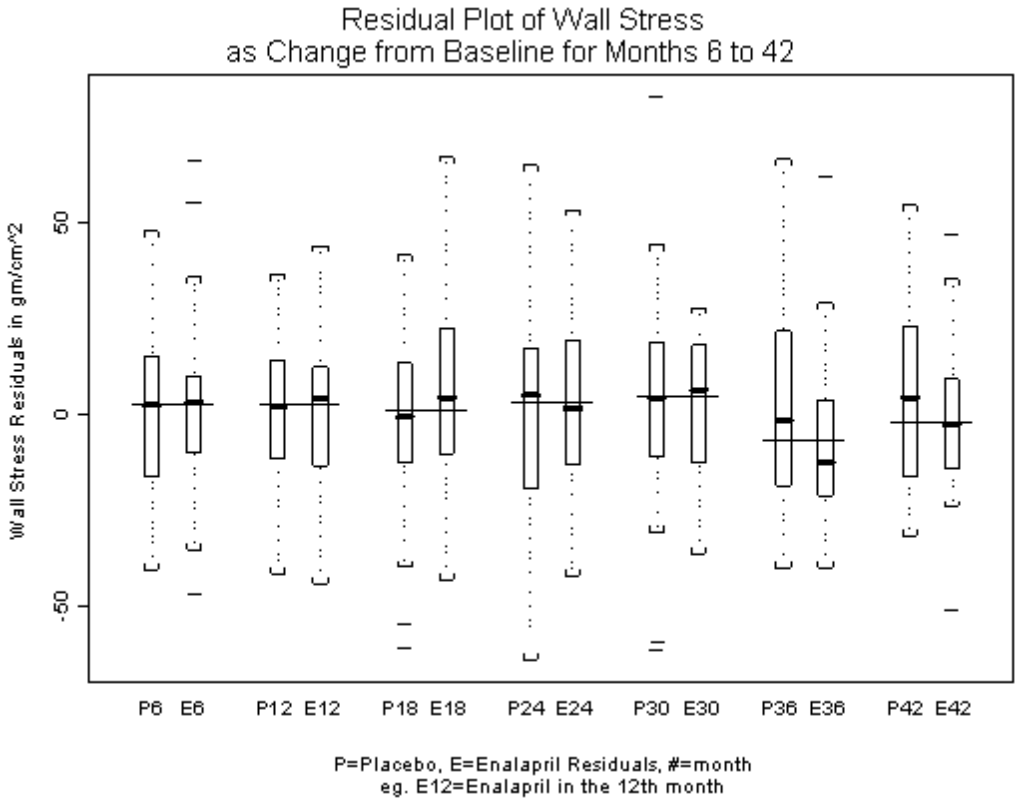
- Significance level for no effect is the same.

- The estimated effect $\widehat{\beta} = -7.11$ at full compliance (dose 1) is about 8% larger than the estimated typical effect $\widehat{\tau} = -6.57$.

20 Model Checks

- Plots of $Y_{sik} - \widehat{\beta}\, D_{sik}$ which should be similar for $Z_i = 1$ (enalapril) and $Z_i = 0$ (placebo) at each time k.

- Also, formal tests based on the T_{sk}'s find little evidence of pattern.


Residual Plot of Wall Stress
as Change from Baseline for Months 6 to 42
Wall Stress Residuals in gm/cm^2
50
0
-50
P6 E6 P12 E12 P18 E18 P24 E24 P30 E30 P36 E36 P42 E42
P=Placebo, E=Enalapril Residuals, #=month
eg. E12=Enalapril in the 12th month

21 Concluding Remarks about Example 1

- IV permits strict adherence to the logic of a randomized trial — randomization forms the basis for inference, and there is no assumption that patients who comply resemble those who do not.

- Also IV agrees exactly with the usual standard — the intent-to-treat analysis — about whether it is plausible that the treatment is without effect.

- However, IV entertains hypotheses in which the treatment effect is present only to the extent that the drug is consumed.

- The basic assumptions of IV — randomness of the instrument, and the exclusion restriction — are plausible here, and the instrument is strong.

22 Why Use Weak Instruments?

- A weak instrument has only a slight impact on the treatment.

- In economics, agents often make choices with a view to achieving objectives. Large biases due to self-selection are anticipated.

- Economists are eager to find rare bits of randomness — even very small bits — to distinguish treatment effects from self-selection.

- Often leads to weak instruments.

23 Angrist & Krueger's (1991) Study

- A very clever example, but one that prompted a critical response in the literature.

- The question concerns the increase in earnings caused by an additional year of schooling.

- Can't simple compare earnings of those with more schooling to earnings of those with fewer years of schooling. Biases due to cognitive abilities, ambition, parental encouragement, financial resources.

- In US, school years begin in September for all children. However, a child may drop out of high school when a certain birthday is reached.

- Quite apart from the child's preferences, compulsory education laws require different children to remain in school for slightly different periods of time.

24 Is there an instrument?

- Angrist and Krueger used quarter-of-birth as an instrument.

- To be an instrument, quarter of birth has to matter for earnings only indirectly through years of schooling. (Plausible, hardly certain.)

- AK demonstrate a saw-tooth pattern in years of education based on quarter-of-birth, consistent with some children dropping out as soon as the law allows. (Perhaps 25% of dropouts.)

- Very weak — adds at most a fraction of a year of schooling to a fraction of students.

- AK write: "Because one's birthday is unlikely to be correlated with personal attributes other than age at school entry, season of birth generates exogenous variation in education..."

25 The AK Data

- AK used microdata from the US Censuses of 1960, 1970, and 1980, with sample sizes $> 200,000$ from each. Obtained data from 50 states and DC about compulsory education laws. Male earnings.

- Typically, children start school in the calender year that they turn 6, so a children born in the fourth quarter ($Z_{si} = 1$) are the younger than others ($Z_{si} = 0$) in the class, so a requirement to stay until a particular birthday requires such children to stay in school longer. 24.5% of men were born in the 4^{th} quarter.

- AK demonstrate that children born in the fourth quarter of the year attend school about $\frac{1}{10}$ of a year longer on average than other children (0.092 years with se 0.013 years). Also, they earn about $\frac{2}{3}$ of 1% more on average (0.0068% with se 0027%).

26 AK Analyses

- AK used quarter of birth as an instrument (Z_{si}) to model log-earnings (Y_{si}) in the census year in terms of years of education (D_{si}).

- Used two-stage least squares or the Wald estimator, obtaining an estimate of 7.4% greater earnings per year of schooling, with standard error 2.8%.

- Wanted to control for 'year of birth' and 'state'. Introduced year and state indicators as covariates and also many instruments formed by interacting quarter of birth with year and state. Obtaining an estimate of 7.4% greater earnings per year of schooling with standard error 0.8%.

- Appears to be the same point estimate, with much greater precision. But alas, ...

27 Analyses, continued

- Bound, Jaeger, Baker (1995) worried about the second analysis. They generated completely random "quarter-of-births" and repeated the second analysis, obtaining very similar results.

- Since then, various papers have shown 2-stage least squares can give inappropriate answers, with inappropriately narrow confidence intervals when the instrument is weak.

- Two problems:

1. Weak instruments may provide little information. (A disappointing feature of some data.)

2. 2-stage least squares suggests there is more information than there is. (A failure of methodology.)

28 Issues

- With weak identification (or no identification), a confidence interval of the form $\widehat{\beta} \pm 2\, \widehat{se}\left(\widehat{\beta}\right)$ may perform very erratically. True of 2SLS.

- First principles: confidence sets formed by inverting hypothesis tests of $H_0 : \beta = \beta_0$ have the possibility of better performance because, with little information, they reject few if any hypotheses, maintaining coverage.

- There is a sense in which the only distribution free confidence intervals for β are formed by inverting a permutation test applied to $Y_{si} - \beta_0\, D_{si}$.

- These intervals may be infinitely long in the case of no identification.

- These intervals may be empty.

29 Logic Again

- Each person is conceived to have a log earnings y_{Tsi} and a years of education d_{Tsi} if born in the fourth quarter ($Z_{si} = 1$), and a log earnings y_{Csi} and years of education d_{Csi} if born in another quarter ($Z_{si} = 0$), with effect proportional to dose, as before:

$$y_{Tsi} - y_{Csi} = \beta\,(d_{Tsi} - d_{Csi})$$

- Key assumption beyond this hypothesis: quarter of birth (Z_{si}) is 'random'.

- If $H_0 : \beta = \beta_0$ were true, then $Y_{si} - \beta_0\, D_{si} = y_{Tsi} - \beta\, d_{Tsi} = y_{Csi} - \beta\, d_{Csi} = a_{si}$.

- Test $H_0 : \beta = \beta_0$ by applying a randomization test to $Y_{si} - \beta_0\, D_{si}$. (Wilcoxon rank sum or permute observations.) Invert for HL estimates and confidence intervals.

30 Outline of Analyses

- Compare 2SLS, rank sum test, permuted observations test, *without* strata. (Fairly similar.)

- Compare 2SLS, rank sum test, permuted observations test, *with* many $year - of - birth \times state$ strata. (Very different.)

- Repeat the Bound, Jaeger, Baker (1995) analysis with randomly generated 'quarters-of-birth,' which imply there is no identification. (Randomization tests figure out that identification is lacking; 2SLS gets it wrong.)

- A small simulation.

Comparison of IV Estimates Without Covariates in the Quarter of Birth Data.

Procedure	$\widehat{\beta}$	95% CI	"se"
2SLS	.074	[.019, .129]	.028
Permute Log Earnings	.073	[.017, .132]	.029
Permute Ranks	.058	[.014, .102]	.023

"se" for permutation tests is the length of the 95% CI divided by 2×1.96.

Comparison of IV Estimates With Year and State Covariates in the Quarter of Birth Data.

Procedure	$\widehat{\beta}$	95% CI	"se"
2SLS	.074	[.058, .090]	.008
Permute Log Earnings	.077	[.036, .139]	.026
Permute Ranks	.067	[−.015, .162]	.045

"se" for permutation tests is the length of the 95% CI divided by 2×1.96.

Repeating the Bound, Jaeger, Baker (1995) Experiment: Noise Instruments

Comparison of IV Estimates with Uninformative Data: Permutation Methods Reveal That the Data Contain No Information, But TSLS is Misleading

Without Covariates	95% Interval
2SLS	$[-.109, .648]$
Permute Log Earnings	Includes $[-1, 1]$
Permute Ranks	Includes $[-1, 1]$
With State and Year Covariates	**95% Interval**
2SLS	$[.042, .078]$
Permute Log Earnings	Includes $[-1, 1]$
Permute Ranks	Includes $[-1, 1]$

$[-1, 1]$ for β is extremely long and entirely uninformative: if $\beta = 1$, then 8 years of additional education for a PhD would multiply a minimum wage salary of perhaps $\$10,000$ per year by a factor of $e^{\beta 8} = 2,980$ to yield $\$30$ million per year; similarly, $\beta = -1$ would reduce $\$10,000$ to $\$183$ per year as a consequence of 4 years of college.

31 Small Simulation

- Two simultaneous equation model:

$$R = r_C + \beta D,$$

$$D = \gamma z + \nu,$$

  where (ν, r_C) may be dependent but are independent of the instrument z, but because of the potential correlation between r_C and ν, the potential response under control, r_C, is not necessarily independent of the dose D, as it would be if doses had been randomly assigned. (For this model, 2SLS = limited-information-maximum likelihood.)

- Four situations, sample size 40, with 10,000 replicates each.

- In all four, the instruments z are normally distributed with zero mean and unit variance.

$$R = r_C + \beta D,\ D = \gamma z + \nu,$$

1. *Strong instrument, thin tails*, $\beta = 1$, $\gamma = 1$. $r_C = \rho \cdot \nu + \sqrt{1 - \rho^2} \cdot \omega$ where ν is N(0,1) and ω is an independent N(0,1), and $\rho = 0.5$.

2. *Strong instrument, thick tails for the response*. $r_C = \rho \cdot \nu + \sqrt{1 - \rho^2} \cdot \omega$ where ν is N(0,1) and ω has a t–distribution with 2 df, and $\rho = 0.5$.

3. *Strong instrument, thick tails for the dose*. $\nu = \rho \cdot r_C + \sqrt{1 - \rho^2} \cdot \omega$ where r_C is standard normal and ω has a t-distribution with 2 df and $\rho = 0.5$.

4. *Weak instrument, thin tails*. (r_C, v) are bivariate Normal with correlation 0.95, so the instrument contributes only slightly to the correlation of dose and responses. The coefficient γ is changed to 0.229 so that the R^2 in the first stage is only 0.05.

32 Simulation Results

- Level and power for $H_0 : \beta = \beta_0$ for various values of β_0 when in fact $\beta = 1$.

- *Strong instrument and thin tails*: all tests have correct size, and TSLS is slightly more powerful.

- *Response equation is thick tailed*: the randomization-based test using the ranks is much more powerful, with more than three times the power of TSLS when testing $H_0 : \beta = 2$.

- *Dose thick tailed*: the standard test is again slightly more powerful.

- *Weak instrument*: TSLS has the wrong size, rejecting true hypotheses approximately 15% of the time, as is well known. The randomization-based tests have correct size.

Figure 1a: quarter of birth data without covariates (solid is rand/rank, dashed is rand/level, dotted is tsls)

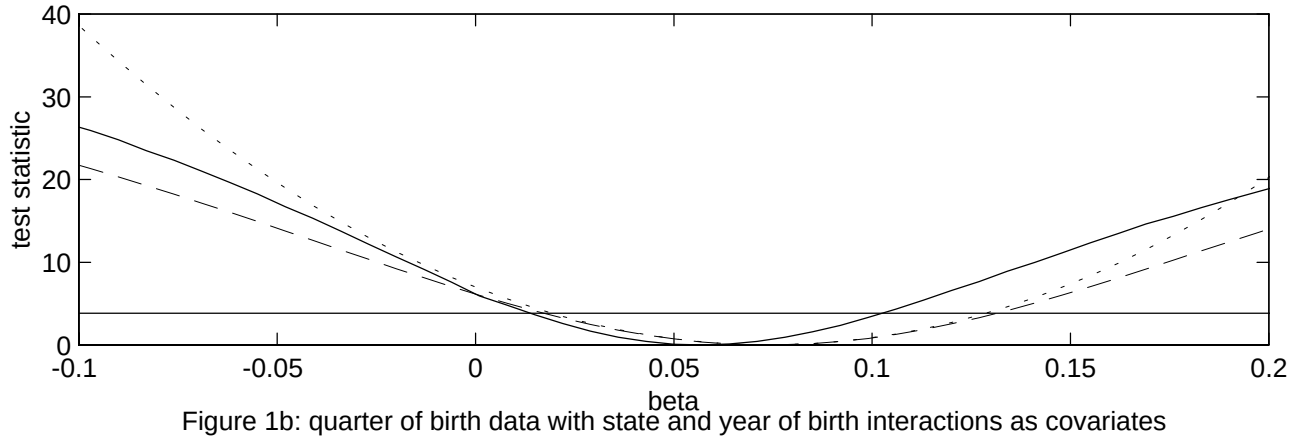


Figure 1b: quarter of birth data with state and year of birth interactions as covariates

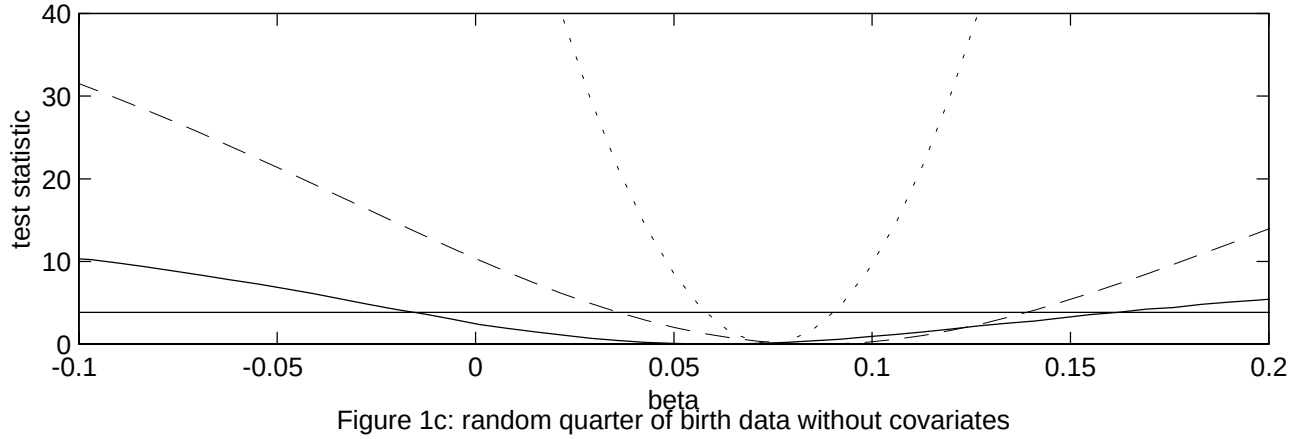


Figure 1c: random quarter of birth data without covariates

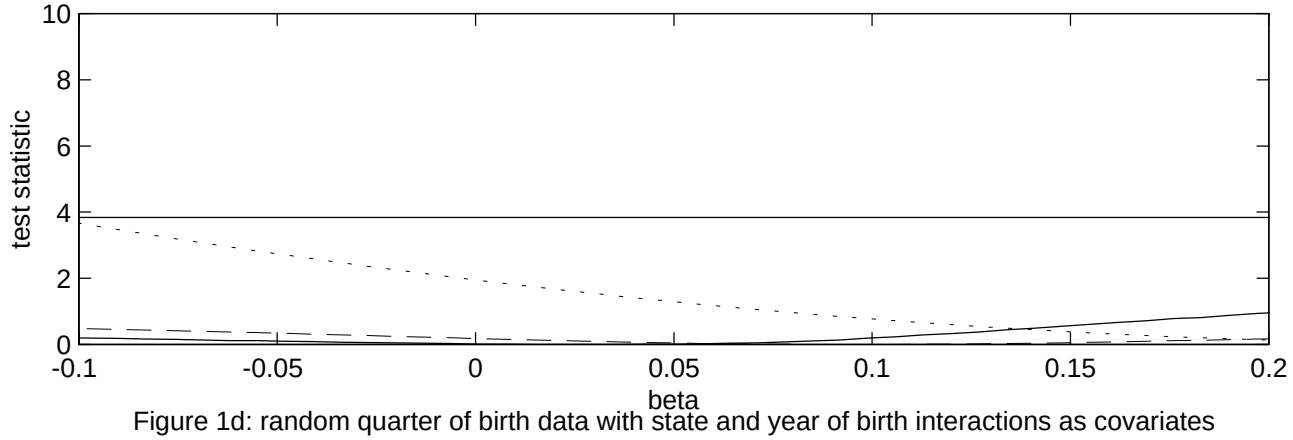


Figure 1d: random quarter of birth data with state and year of birth interactions as covariates

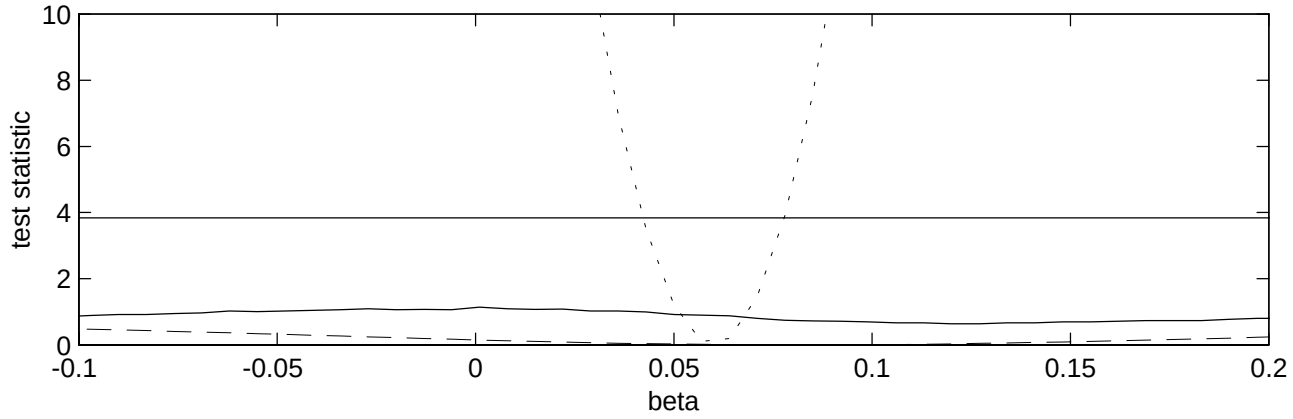


Fig 2a: Power Function Design 1 (solid is rand/rank, dashed is rand/level, dotted is tsls)

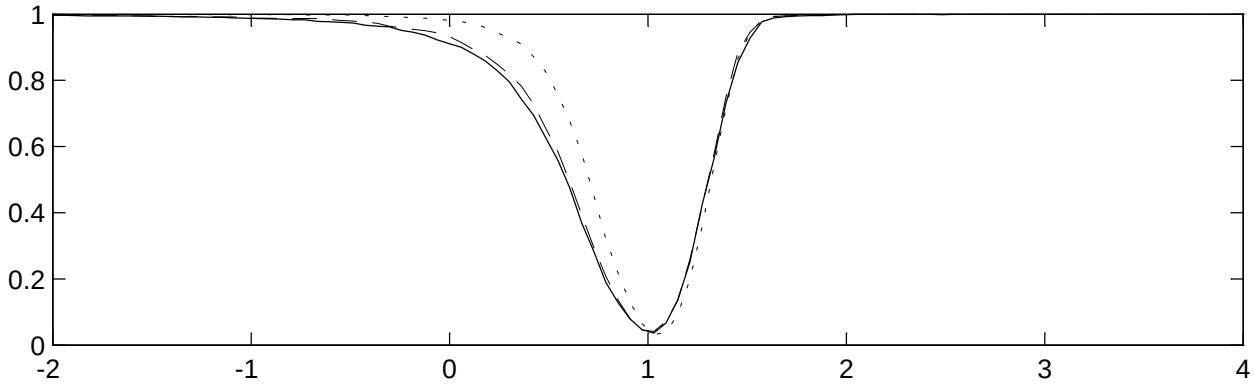


Fig 2b: Power Function Design 2

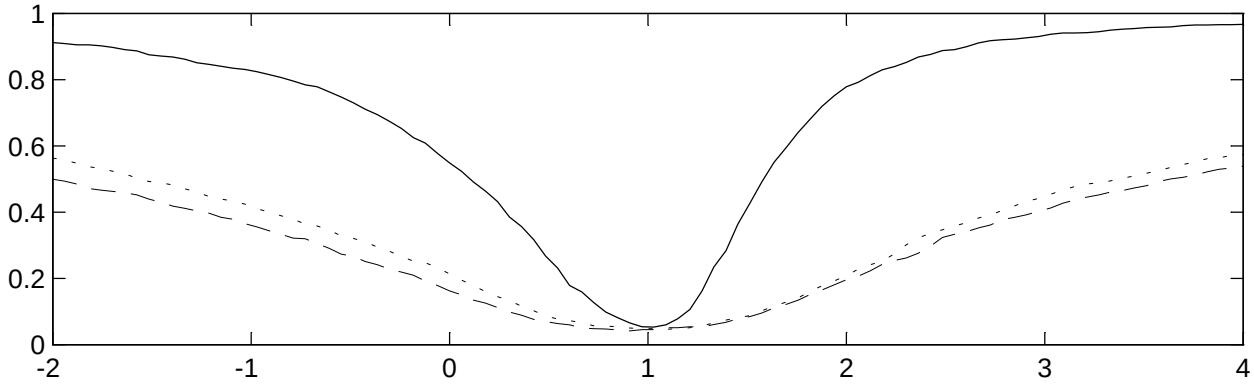


Fig 2c: Power Function Design 3

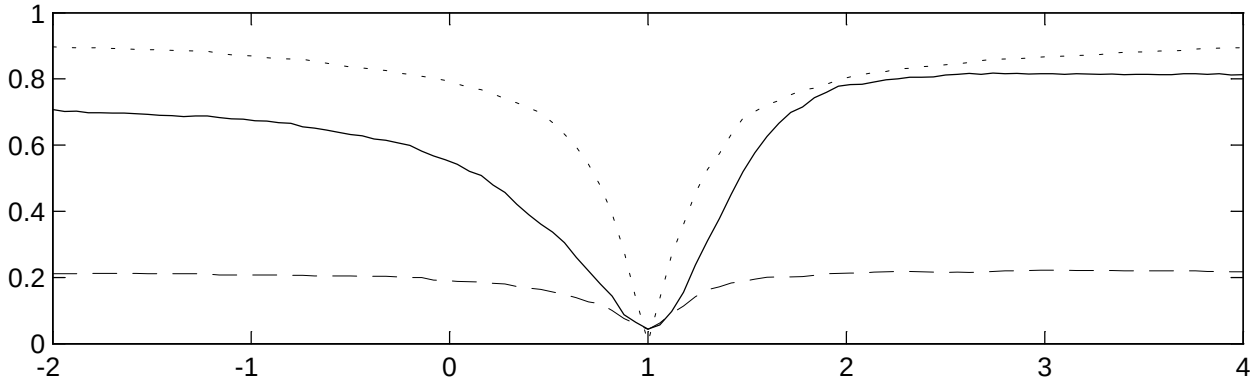
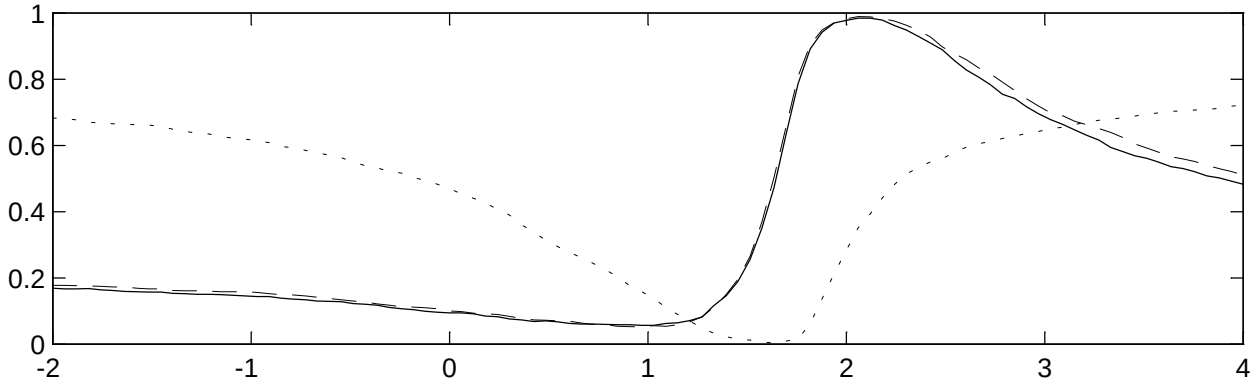


Fig 2d: Power Function Design 4



33 Summary

- If the assigned treatment is used as an instrument for the dose received in a randomized experiment, one may adhere to the tight logic of the randomized trial, while acknowledging that a medication will not have its biological effects if it remains in the bottle.

- With weak instruments, randomization inference yields confidence intervals with correct coverage. If the identification is weak or absent, the interval maintains its coverage by becoming longer, possibly infinite in length.

- Instead of assuming adequate identification, randomization inference permits the data to speak to the issue.

34 E1: Toy Example

- One stratum, $S = 1$, drop the s. Randomize $n = 10$ subjects, $m = 5$ to be encouraged to exercise, $Z_i = 1$, and $n - m = 5$ to no encouragement, $Z_i = 0$.

- Then (d_{Tsi}, d_{Csi}) indicates whether subject i would exercise with and without encouragement. D_i is whether you did exercise, given the treatment to which you were randomly assigned.

- The response (r_{Tsi}, r_{Csi}) is FEV on a suitable scale.

- Effect of exercise is $r_{Ti} - r_{Ci} = \beta\,(d_{Ti} - d_{Ci})$ with $\beta = 5$, so $r_{Ti} - r_{Ci} = 0$ if exercise behavior would not change, and $r_{Ti} - r_{Ci} = 5$ if exercise behavior would increase by encouragement. *Efficacy* of exercise is $\beta = 5$, but *effectiveness* of encouragement is $\beta\,(d_{Ti} - d_{Ci})$.

35 E2: Toy Example, data

i	d_{Ti}	d_{Ci}	r_{Ti}	r_{Ci}	Z_i	D_i	R_i
1	1	1	71	71	1	1	71
2	1	1	68	68	0	1	68
3	1	0	64	59	1	1	64
4	1	0	62	57	0	0	57
5	1	0	59	54	0	0	54
6	1	0	58	53	1	1	58
7	1	0	56	51	1	1	56
8	1	0	56	51	0	0	51
9	0	0	42	42	0	0	42
10	0	0	39	39	1	0	39

Effect is 5: $r_{Ti} - r_{Ci} = 5\,(d_{Ti} - d_{Ci})$

$$D_i = 1 \text{ vs } D_i = 0 :$$

$$\frac{71+68+64+58+56}{5} - \frac{57+54+51+42+39}{5} = 14.8$$

$$Z_i = 1 \text{ vs } Z_i = 0 :$$

$$\frac{71+64+58+56+39}{5} - \frac{68+57+54+51+42}{5} = 3.2$$

36 E3: Randomization Inference

$H_0 : \beta = \beta_0$ is tested using $T = t(\mathbf{Z}, \mathbf{g}) = t(\mathbf{Z}, \mathbf{R} - \beta_0 \mathbf{D})$, say the Wilcoxon rank sum test. Solving for $\widehat{\beta}$ in

$$t\left(\mathbf{Z}, \mathbf{R} - \widehat{\beta}\mathbf{D}\right) = \frac{m(n+1)}{2} = \frac{5(10+1)}{2} = 27.5$$

gives $\widehat{\beta} = 5$.

i	R_i	D_i	Z_i	$R_i - 5D_i$	Rank
1	71	1	1	66	10
2	68	1	0	63	9
3	64	1	1	59	8
4	57	0	0	57	7
5	54	0	0	54	6
6	58	1	1	53	5
7	56	1	1	51	3.5
8	51	0	0	51	3.5
9	42	0	0	42	2
10	39	0	1	39	1

37 T1: All Distribution Free IV Inferences Are Permutation Inferences

- Recall that the adjusted responses were the same under treatment and control:

$$\begin{aligned} Y_{si} - \beta D_{si} &= y_{Tsi} - \beta\, d_{Tsi} \\ &= y_{Csi} - \beta\, d_{Csi} = a_{si}, \text{ say} \end{aligned}$$

- Suppose that, for each stratum s, the adjusted responses a_{si} were iid from a continuous distribution with unknown density $f_s(\cdot)$, so that their joint density is $\prod_{s=1}^{S} \prod_{i=1}^{n_s} f_s(a_{si})$.

- It then follows that the only tests of $H_0 : \beta = \beta_0$ are permutation or randomization tests (ie tests that permute the a_{si}'s relative to the Z_{si}'s, rejecting at level α for $100\alpha\%$ of these permutations).

38 T2: All Distribution Free IV Inferences Are Permutation Inferences

- The proof is essentially the same as the proof in Lehmann (1959, §5.7, Prop. 3). Details: Imbens & Rosenbaum manuscript.

- Formally, Lehmann discusses the case of an additive treatment effect, $y_{Tsi} = y_{Csi} + \tau$, in which case the adjusted responses a_{si} are $Y_{si} - \tau Z_{si}$.

- However, the proof concerns behavior of the test when the null hypothesis is true, and in both cases, the adjusted responses have continuous distribution $\prod_{s=1}^{S} \prod_{i=1}^{n_s} f_s(a_{si})$ with the f_s unspecified, and the remainder of the proof is the same.

39 T3: Idea of Proof Given by Lehmann

- Fix the Z_{si}'s and permute the a_{si}. (Only relative order matters.)

- Test function $\phi(\mathbf{a}) = 1$ for rejection, $= 0$ for acceptance. Consider tests with $E\{\phi(\mathbf{a})\} = \alpha$ for all f_s's.

- The order statistics, say $\overrightarrow{\mathbf{a}}$, of $\mathbf{a}$, within strata are a complete sufficient statistic for the f_s's. The conditional probability of rejection given the order statistic is $E\left\{\phi(\mathbf{a}) \,|\, \overrightarrow{\mathbf{a}}\right\}$.

- Obviously

$$\alpha = E\{\phi(\mathbf{a})\} = E\left[E\left\{\phi(\mathbf{a}) \,|\, \overrightarrow{\mathbf{a}}\right\}\right]$$

but by complete sufficiency, this implies

$$\alpha = E\left\{\phi\left(\mathbf{a}\right)|\overrightarrow{\mathbf{a}}\right\} \text{ for almost all } \overrightarrow{\mathbf{a}}$$

wwwww.