Some new insights on the Fisher randomization test

Texas A&M Data Science Seminar

Tirthankar Dasgupta

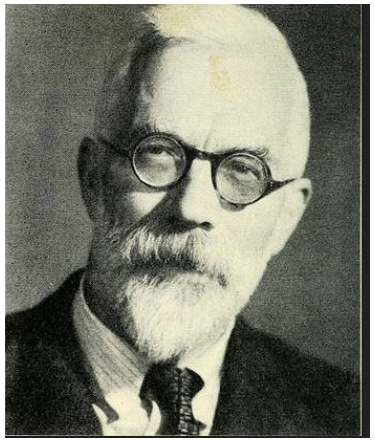
Department of Statistics

Rutgers University

November 1, 2021

The randomization-based perspective of experimental design

- Long served as the foundation of experimental design.
- Seminal work by two stalwarts (Neyman 1923, Fisher 1925, 1935).
- Connection with survey sampling.
- Died down in the later half of the twentieth century – lack of computational resources?



Motivating example: A simplified education experiment

- Examining the impact of school-wide performance bonus scheme for teachers on performance of school
- Experimental units: 12 schools
- Treatment: Implement performance bonus scheme, Control: Do not implement
- Response: Year-end performance score (1-100)
- We will review some basic designs for this experiment and their popular analysis methods that we are no doubt all familiar with, and then see how randomization-based analysis can replace them.

Completely randomized design (CRD)

- Experimental unit: 12 schools
- Completely randomized allocation
  - 6 receive treatment (T=1)
  - 6 receive control (T=0)
- Response: Performance score (1-100)

Unit	1	2	3	4	5	6	7	8	9	10	11	12
T	1	0	1	0	0	1	0	0	1	1	1	0
Score	66.85	70.52	68.53	57.34	66.89	68.53	59.52	59.22	66.02	72.58	64.34	58.40

- How to analyze the data?

Two-sample t-test in R

Unit	1	2	3	4	5	6	7	8	9	10	11	12
T	1	0	1	0	0	1	0	0	1	1	1	0
Score	66.85	70.52	68.53	57.34	66.89	68.53	59.52	59.22	66.02	72.58	64.34	58.40

```
> t.test(y1,y0)

        Welch Two Sample t-test

data:  y1 and y0
t = 2.3463, df = 7.5603, p-value = 0.04874
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  0.04155153 11.61178180
sample estimates:
mean of x mean of y
 67.80833  61.98167
```

Randomized block design (RBD)

- Experimental unit: 12 schools
- Covariate information (number of students) available
- Assumption: Performance of school is strongly dependent on the total number of students (covariate X)
- Modify the design to prevent confounding of treatment effects with X
- Create two blocks: schools with less than 1000 students (blue cells), and schools with greater than or equal to 1000 students (yellow cells)

Unit	1	2	3	4	5	6	7	8	9	10	11	12
x	1165	748	1010	1157	482	917	1108	823	1293	566	1089	689

Randomized block design (contd.)

- In each block, assign three units to treatment and three units to control using a completely randomized assignment

	BLOCK 1						BLOCK 2					
Unit	1	3	4	7	9	11	2	5	6	8	10	12
x	1165	1010	1157	1108	1293	1089	748	482	917	823	566	689
T	1	1	0	1	0	0	0	0	1	0	1	1
Y	66.84	69.66	58.34	67.17	55.76	59.90	64.74	64.01	71.79	68.70	73.26	68.63

Analysis of RBD

```
Data <- M[,3:6]
```

```
 x   B    T  yobs
1165 1    1 66.84
 748 2    0 64.74
1010 1    1 69.66
1157 1    0 58.34
 482 2    0 64.01
 917 2    1 71.79
1108 1    1 67.17
 823 2    0 68.70
1293 1    0 55.76
 566 2    1 73.26
1089 1    0 59.90
 689 2    1 68.63
```

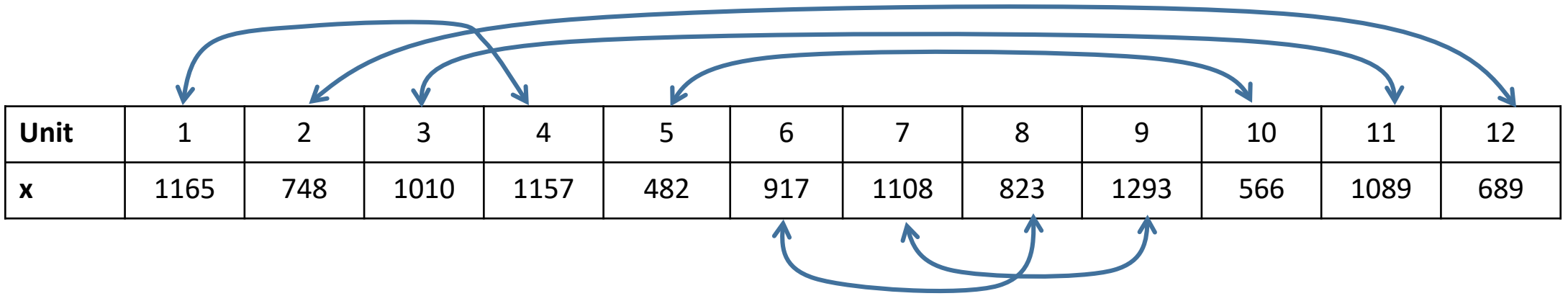
```
> rbd <- lm(yobs ~ factor(B) + factor(T))
> anova(rbd)
Analysis of Variance Table

Response: yobs
             Df  Sum Sq Mean Sq F value    Pr(>F)
factor(B)     1  93.298  93.298  16.000 0.0031106 **
factor(T)     1 175.568 175.568  30.108 0.0003866 ***
Residuals     9  52.481   5.831
---
```

Matched-pair design

- Experimental unit: 12 schools
- Assumption: Performance of school is very strongly dependent on the total number of students – even a difference of 200-300 students may make a major difference.
- Solution: Create more blocks; extreme case: pair the 12 schools based on number of students
- Within each pair one receives treatment and one receives control with equal probability

Matched-pair design and outcomes



Unit	1	2	3	4	5	6	7	8	9	10	11	12
x	1165	748	1010	1157	482	917	1108	823	1293	566	1089	689

Unit	1	4	2	12	3	11	5	10	6	8	7	9
x	1165	1157	748	689	1010	1089	482	566	917	823	1108	1293
T	1	0	1	0	0	1	1	0	1	0	1	0
Y	68.35	58.43	72.52	63.11	59.90	69.11	75.18	64.34	70.83	61.77	68.92	57.07

Analysis of matched-pair design

```
> matched_data
   Bmatch    x    T  yobs
1       1 1165    1 68.35
4       1 1157    0 58.43
2       2  748    1 72.52
12      2  689    0 63.11
3       3 1010    0 59.90
11      3 1089    1 69.11
5       4  482    1 75.18
10      4  566    0 64.34
6       5  917    1 70.83
8       5  823    0 61.77
7       6 1108    1 68.92
9       6 1293    0 57.07
```

```
> matched_analysis <- lm(matched_data$yobs ~
factor(matched_data$Bmatch) + factor(matched_data$T))
> anova(matched_analysis)
Analysis of Variance Table

Response: matched_data$yobs
                            Df  Sum Sq Mean Sq F value    Pr(>F)
factor(matched_data$Bmatch)  5  70.690  14.138  23.657  0.001722 **
factor(matched_data$T)       1 302.907 302.907 506.849 3.213e-06 ***
Residuals                    5   2.988   0.598
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The steps in model-based analysis

- Hypothesis testing framework: equality of treatment effects
- Linear model, assumes normality of residuals
- Find a suitable test statistic
- Distribution of a test statistic under the null hypothesis (obtained from the model)
- Reject null if the observed value of the test statistic is very unusual (extreme) with respect to its null (small p-value)

Model-based analysis to model-free analysis

- Hypothesis testing framework: equality of treatment effects (for all units)
- ~~Linear model, assumes normality of residuals~~ – No model, no normality assumption
- ~~Find a suitable test statistic~~ Take ANY reasonable test statistic
- Distribution of a test statistic under the null hypothesis ~~(obtained from the model)~~ – obtained from randomization distribution
- Reject null if the observed value of the test statistic is very unusual (extreme) with respect to its null (small p-value)

Randomization analysis of CRD data: test statistic and its randomization distribution under null

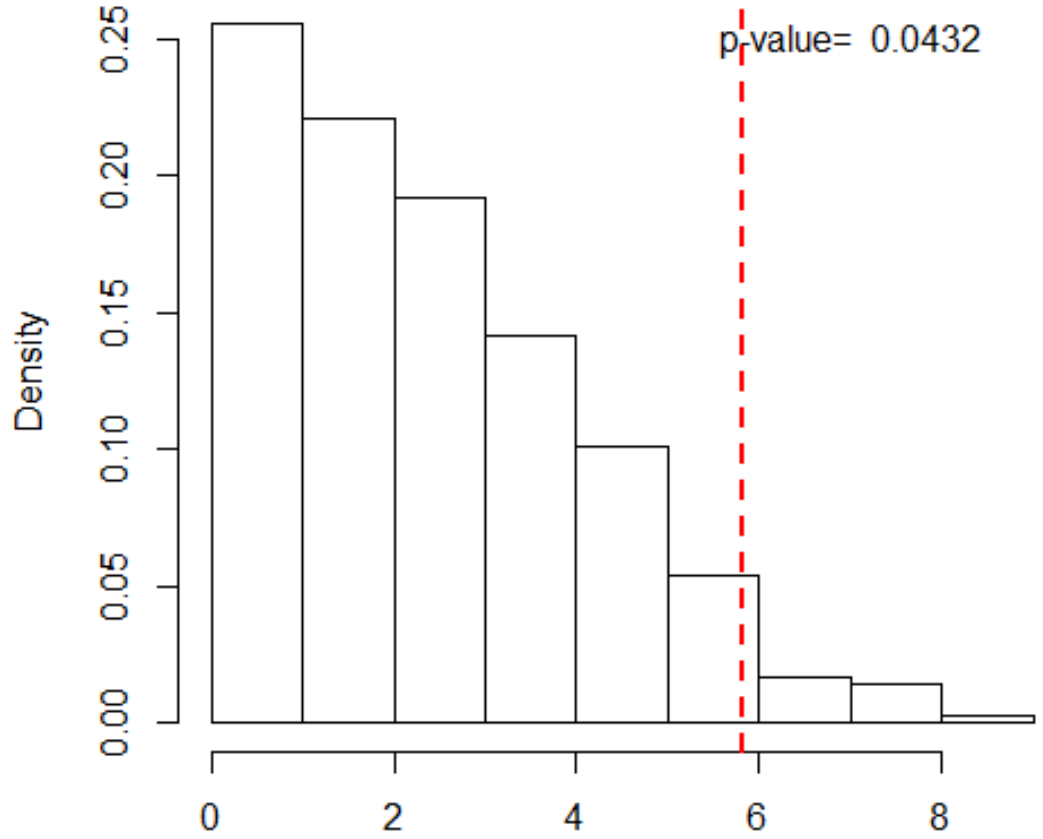
- Consider absolute difference of means in the two treatment groups as the test statistic. Observed value = |67.81 - 61.98| = 5.83.

	Unit	1	2	3	4	5	6	7	8	9	10	11	12	Abs diff of means
	Score	66.85	70.52	68.53	57.34	66.89	68.53	59.52	59.22	66.02	72.58	64.34	58.40	
Observed	**1**	1	0	1	0	0	1	0	0	1	1	1	0	5.83
	2	0	0	0	1	1	1	0	1	0	0	1	1	8.29
	3	1	1	1	1	1	0	0	0	1	0	0	0	7.89
	…..	…..	…..	…..	…..	…..	…..	…..	…..	…..	…..	…..	…..	…..
	924	0	0	1	1	1	0	0	0	0	1	1	1	9.08

12!/(6! 6!) = 924 allowable assignments

Assumption (sharp null): the outcome of each unit would be unchanged under a different assignment (i.e., no treatment effect on ANY unit)

Randomization distribution and p-value: CRD



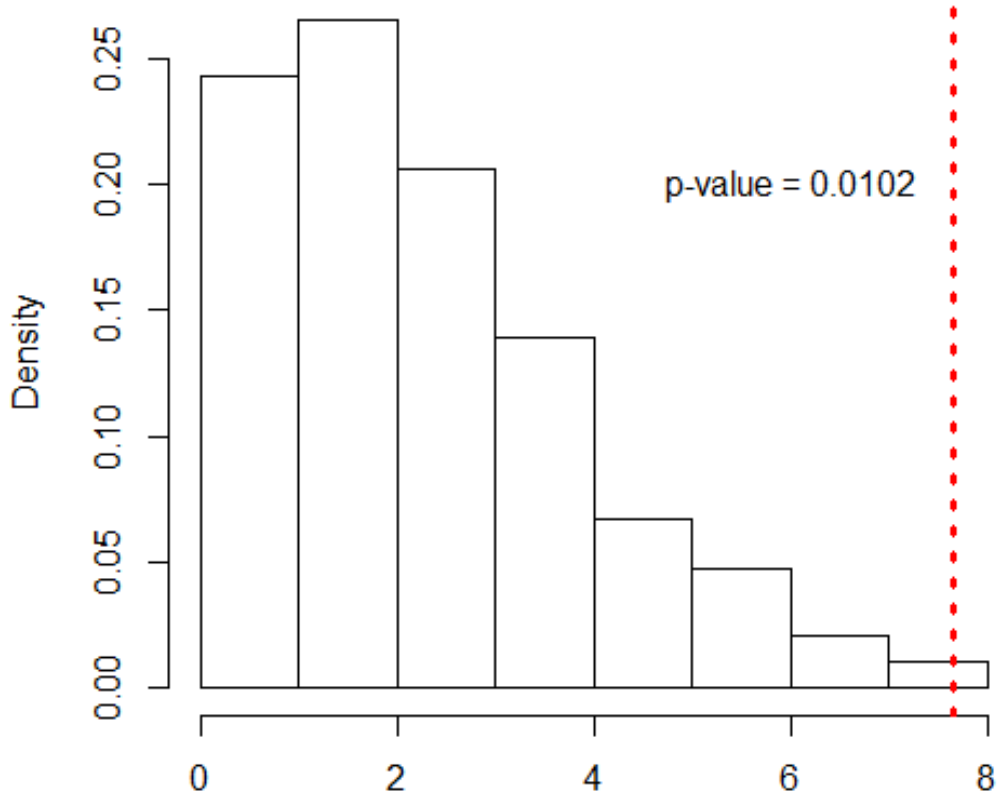
Randomization analysis of RBD data: test statistic and its randomization distribution under null

- Consider absolute difference of means in the two treatment groups as the test statistic. Observed value = 7.65 .

	No	1	3	4	7	9	11	2	5	6	8	10	12	Abs diff of means
Observed	**1**	1	1	0	1	0	0	0	0	1	0	1	1	7.65
	2	1	0	1	1	0	0	0	1	0	1	0	1	3.68
	3	1	0	1	0	0	1	0	1	1	1	0	0	3.87
	…..	**…..**	**…..**	**…..**	**…..**	**…..**	**…..**	**…..**	**…..**	**…..**	**…..**	**…..**	**…..**	**…..**
	400	0	1	1	1	0	0	1	1	0	0	0	1	0.95

6!/(3! 3!) x 6!/(3! 3!) = 400 allowable assignments

Randomization distribution and p-value: RBD



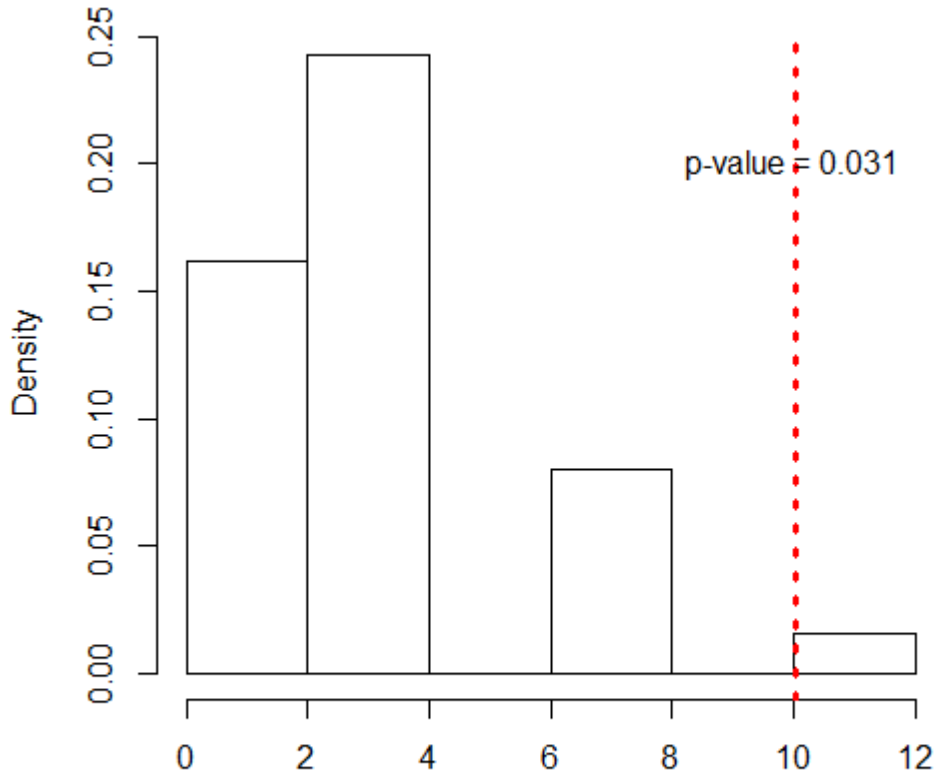
Randomization analysis of Matched-pair data: test statistic and its randomization distribution under null

- Consider absolute difference of means in the two treatment groups as the test statistic. Observed value = 6.47 **.**

	No	1	4	2	12	3	11	5	10	6	8	7	9	Abs diff of means
Observed	1	1	0	1	0	0	1	1	0	1	0	1	0	10.05
	2	1	0	0	1	0	1	1	0	0	1	0	1	0.26
	…..	…..	…..	…..	…..	…..	…..	…..	…..	…..	…..	…..	…..	…..
	64	0	1	1	0	0	1	0	1	0	1	0	1	0.77

2^6= 64 allowable assignments

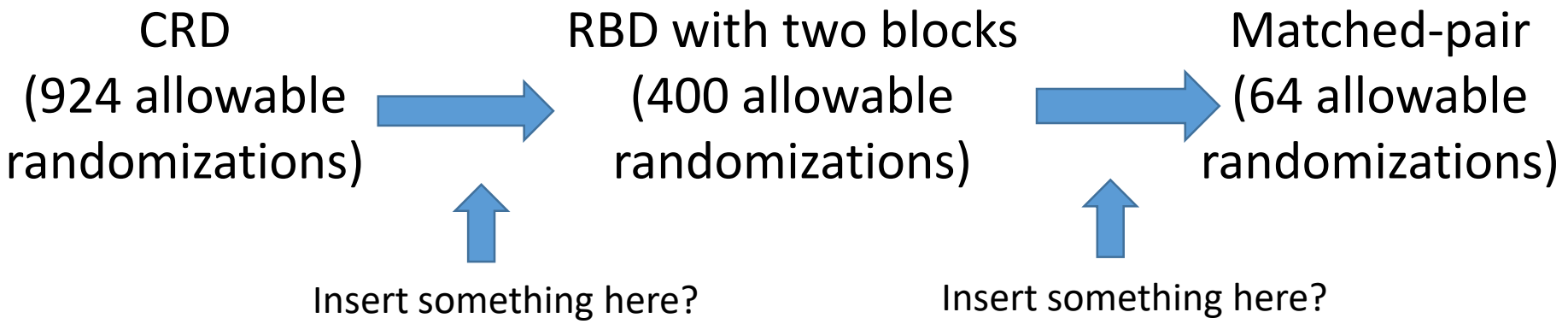
Randomization distribution and p-value: MP



Design and randomization-based analysis of randomized experiments

- Design: Define ALLOWABLE randomizations (prevent confounding with covariates)
  - Bernoulli design: $2^{12} = 4096$
  - CRD: 924
  - RBD with two blocks: 400
  - Matched-pair (RBD with six blocks of size two each): 64
- Analyze using randomization test:
  - Choose test statistic (ANY reasonable statistic works)
  - Calculate observed value of test statistic
  - Generate distribution of test statistic using repeated assignments under allowable randomizations
  - Determine if observed value is “unusual”

Do we need to stick to “standard” designs?



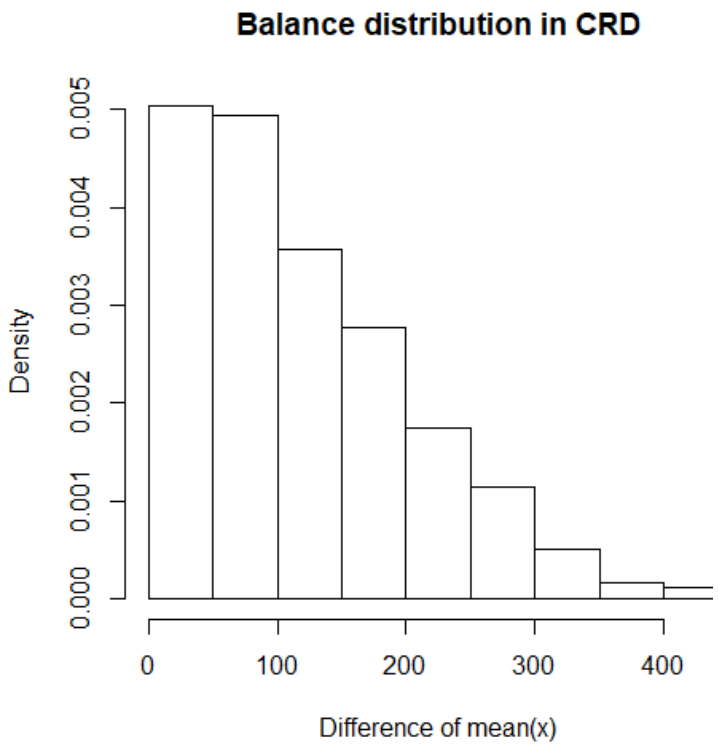
- Define a measure of balance between treated and control groups
- Certain values of the measure indicate balanced groups and are acceptable
- Certain values of the measure indicate lack of balance and are unacceptable

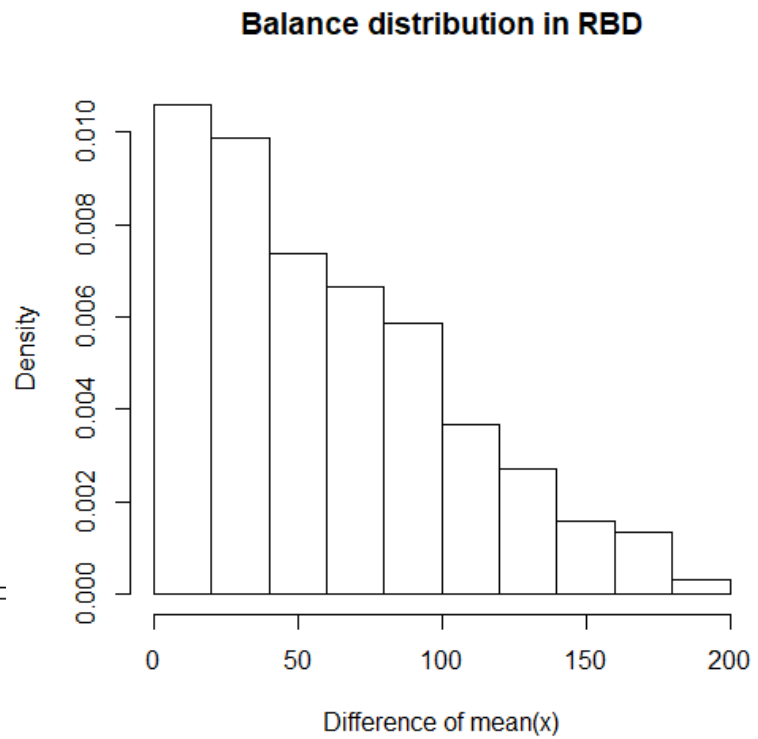
Balance measure for different assignments

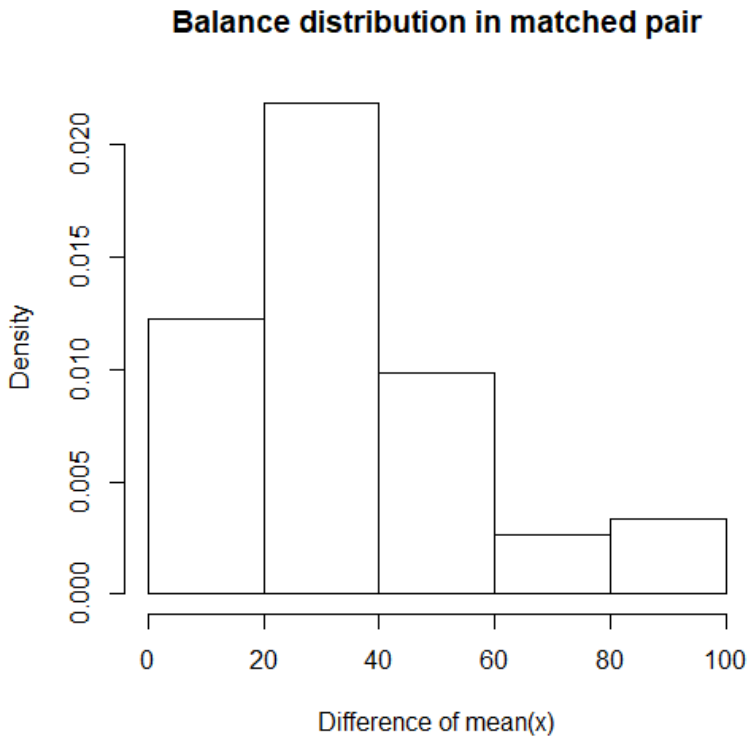
Balance measure: Absolute Difference in x (enrollment) between treatment groups

Unit	1	2	3	4	5	6	7	8	9	10	11	12	Balance
x	1165	748	1010	1157	482	917	1108	823	1293	566	1089	689	
1 (CRD)	1	0	1	0	0	1	0	0	1	1	1	0	352.83
2 (RBD)	1	0	1	0	0	1	1	0	0	1	0	1	22.833
3 (MP)	1	0	1	0	0	1	1	0	1	0	1	0	4.833
…..	…..	…..	…..	…..	…..	…..	…..	…..	…..	…..	…..	…..	…..	…..
924	0	1	0	1	0	0	1	0	1	0	1	1	186.83

How the balance improves from CRD to MP







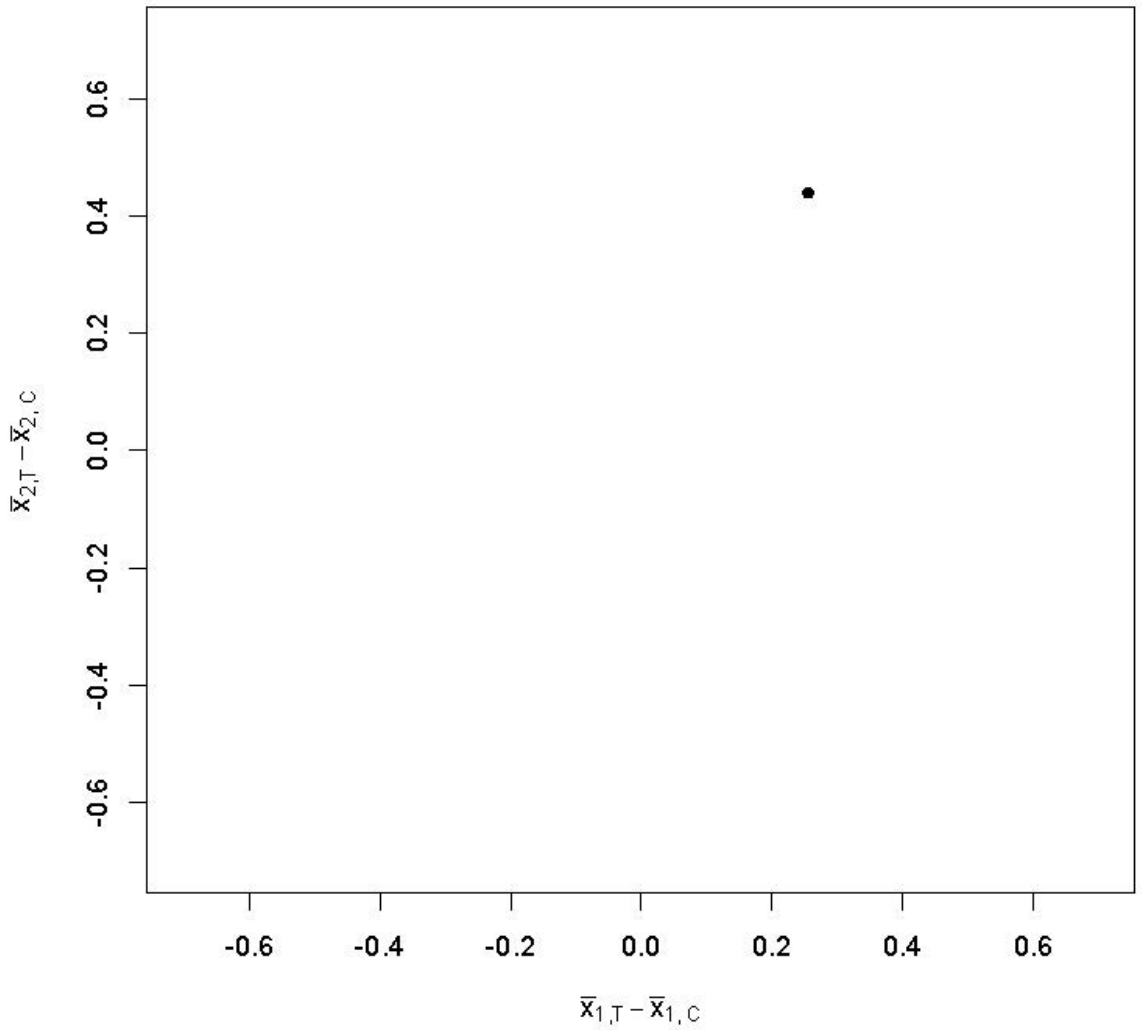
We can choose any reasonable cut-off for the balance measure

- We can declare all assignment vectors that yield balance measures not exceeding 300 as acceptable
  - Stricter than CRD but less strict than RBD
- Or we can declare all assignment vectors that yield balance measures not exceeding 150 as acceptable
  - Stricter than RBD but less strict than MP
- The randomization test for the above two designs can be performed in exactly the same manner as before
- Make sure that “an allowable” randomization is defined before you design the experiment and the same rule is followed during analysis while generating the randomization distribution of the statistic

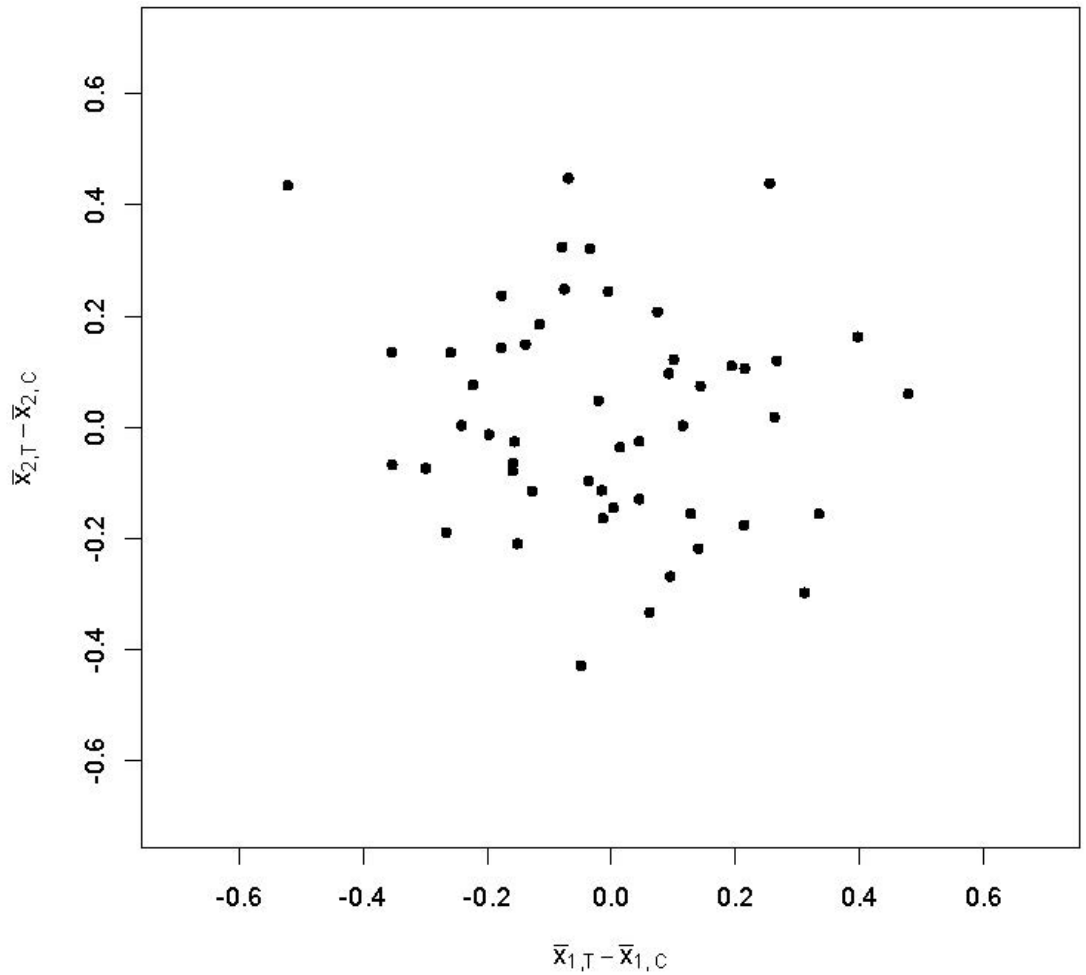
When blocking is no longer intuitive: multiple covariates

- Large number of covariates associated with each experimental unit
  - previous year's performance score,
  - total number of students,
  - race variables (proportion of white, black, Asian, Native American and Latino students),
  - proportion of female students
  - enrollment rate
  - poverty rate in neighborhood
  - ….. and many more

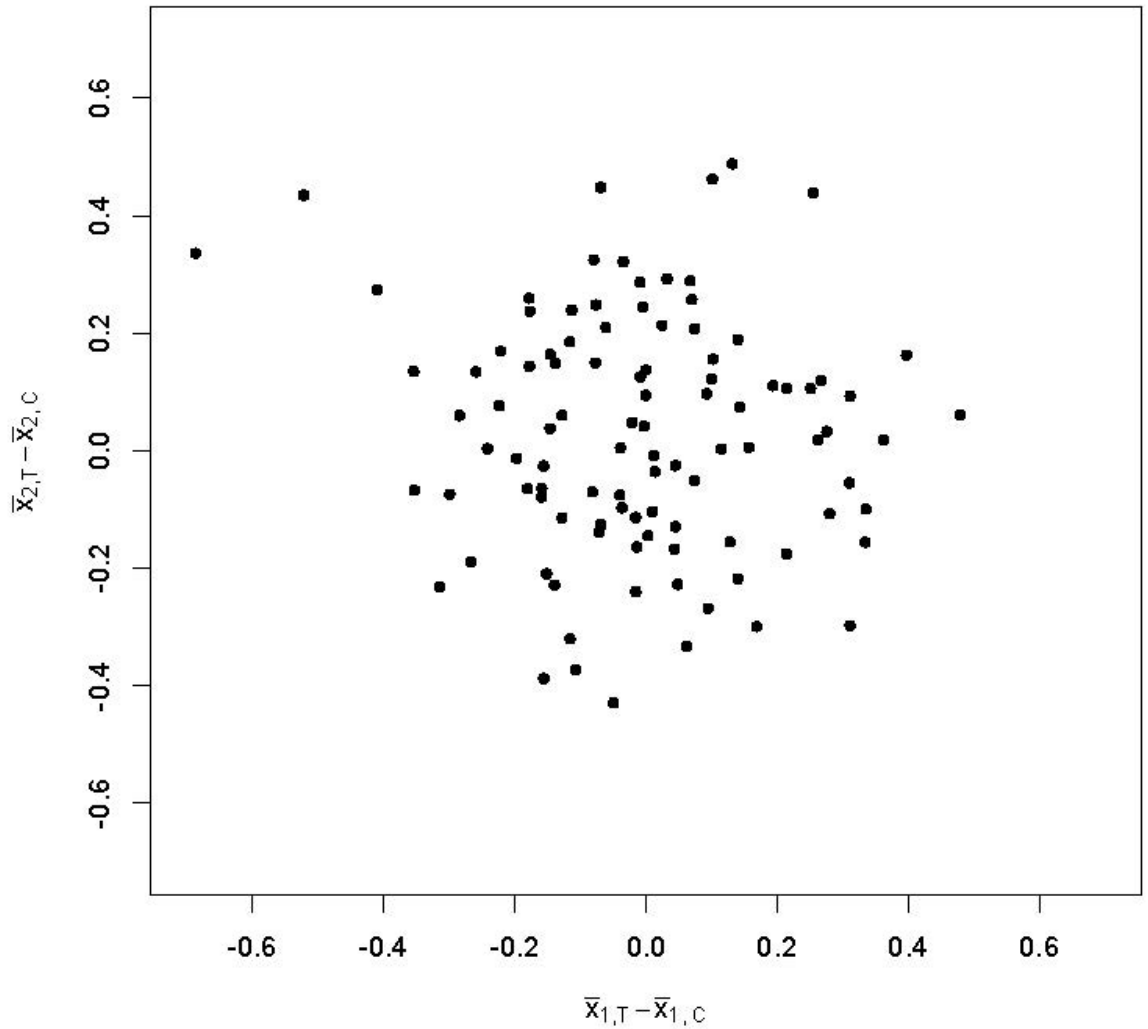
Visualization for two continuous covariates



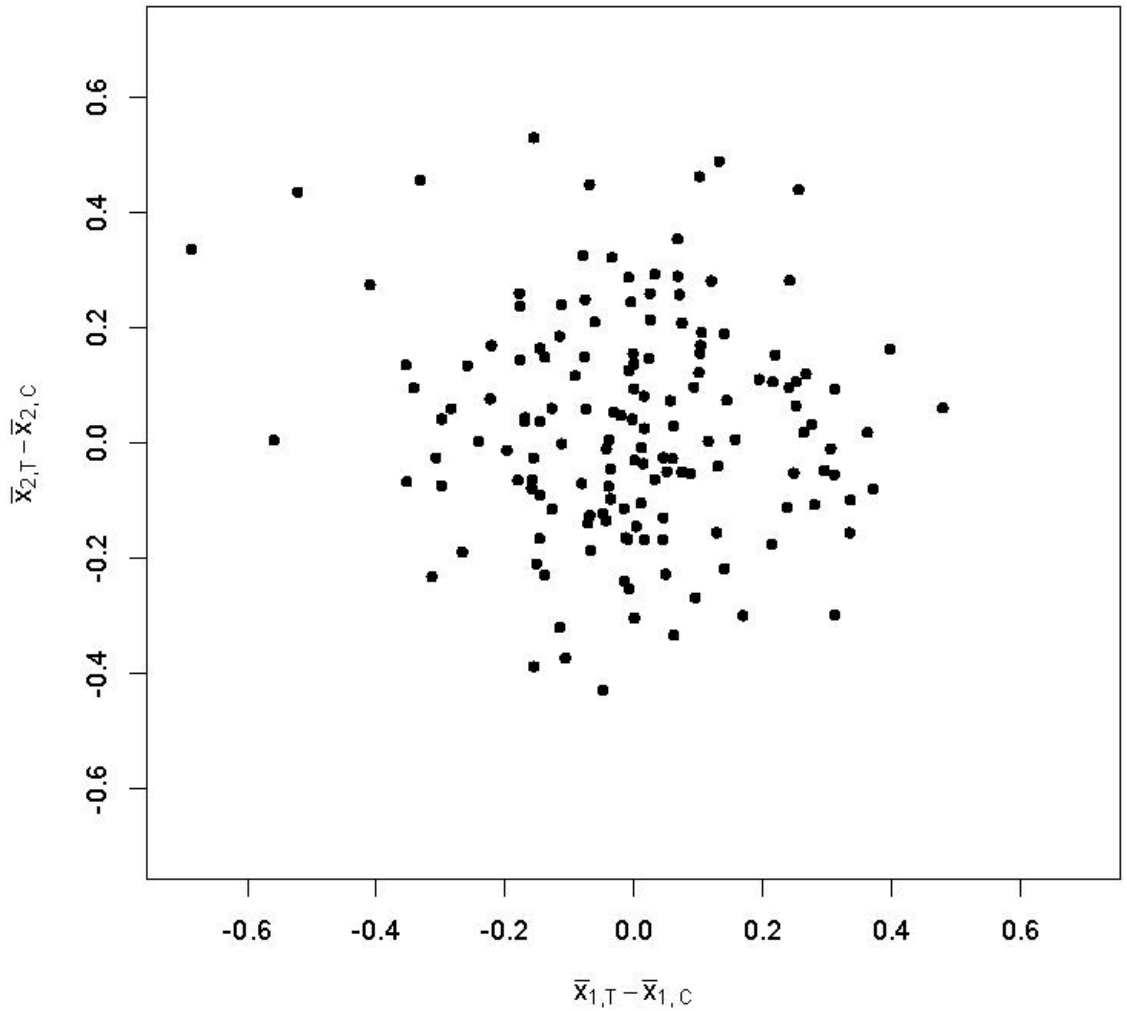
Visualization for two continuous covariates



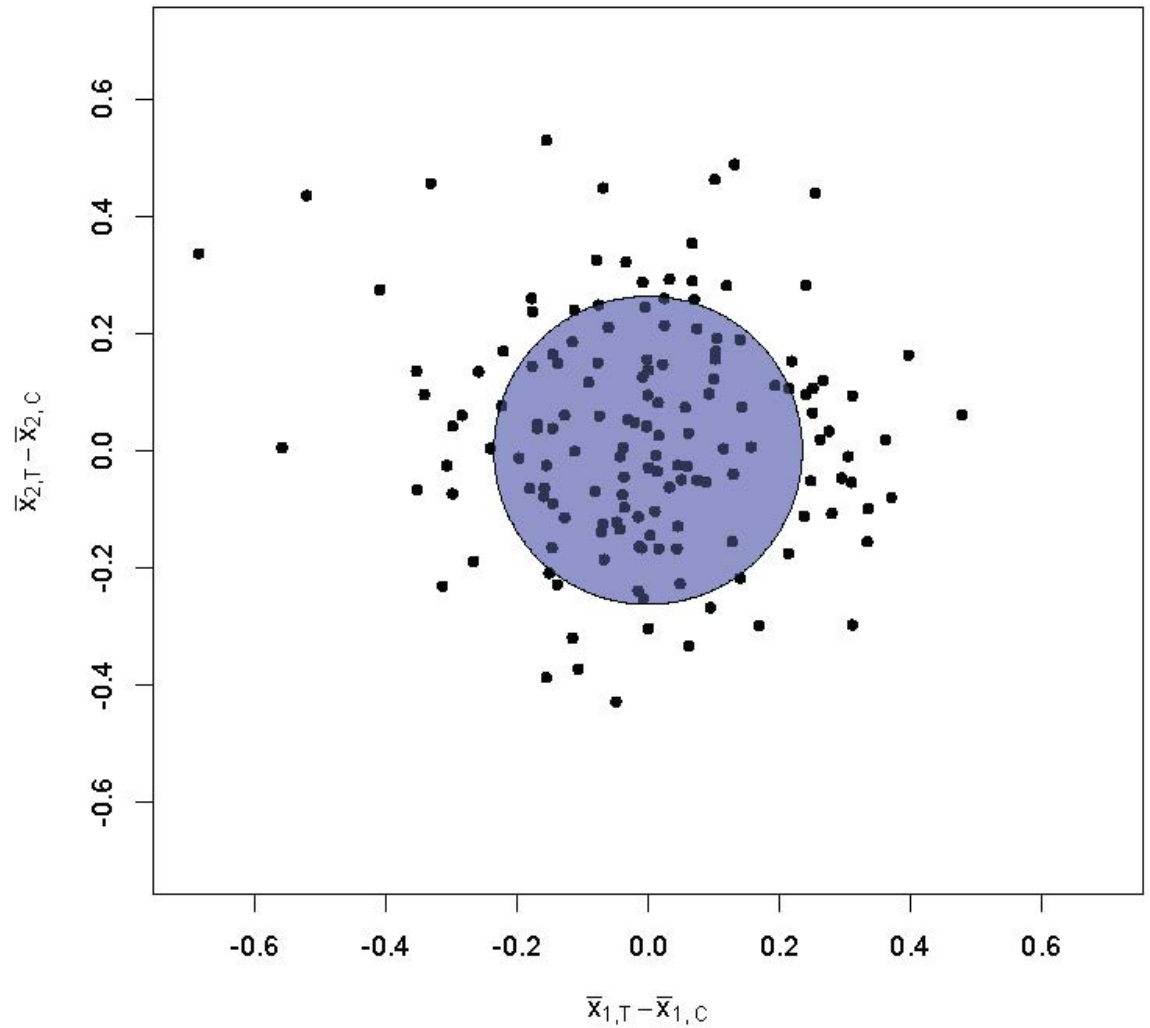
Visualization for two continuous covariates



Visualization for two continuous covariates



Visualization for two continuous covariates



Criterion for randomization

- Mahalanobis distance *M* (a multivariate distance between group mean vectors)
- Acceptance criterion: $M \leq a$
- Here *a* is a pre-determined constant
- Trade-off between throwing away randomizations and balancing groups

Reducing variance of average covariate difference between groups

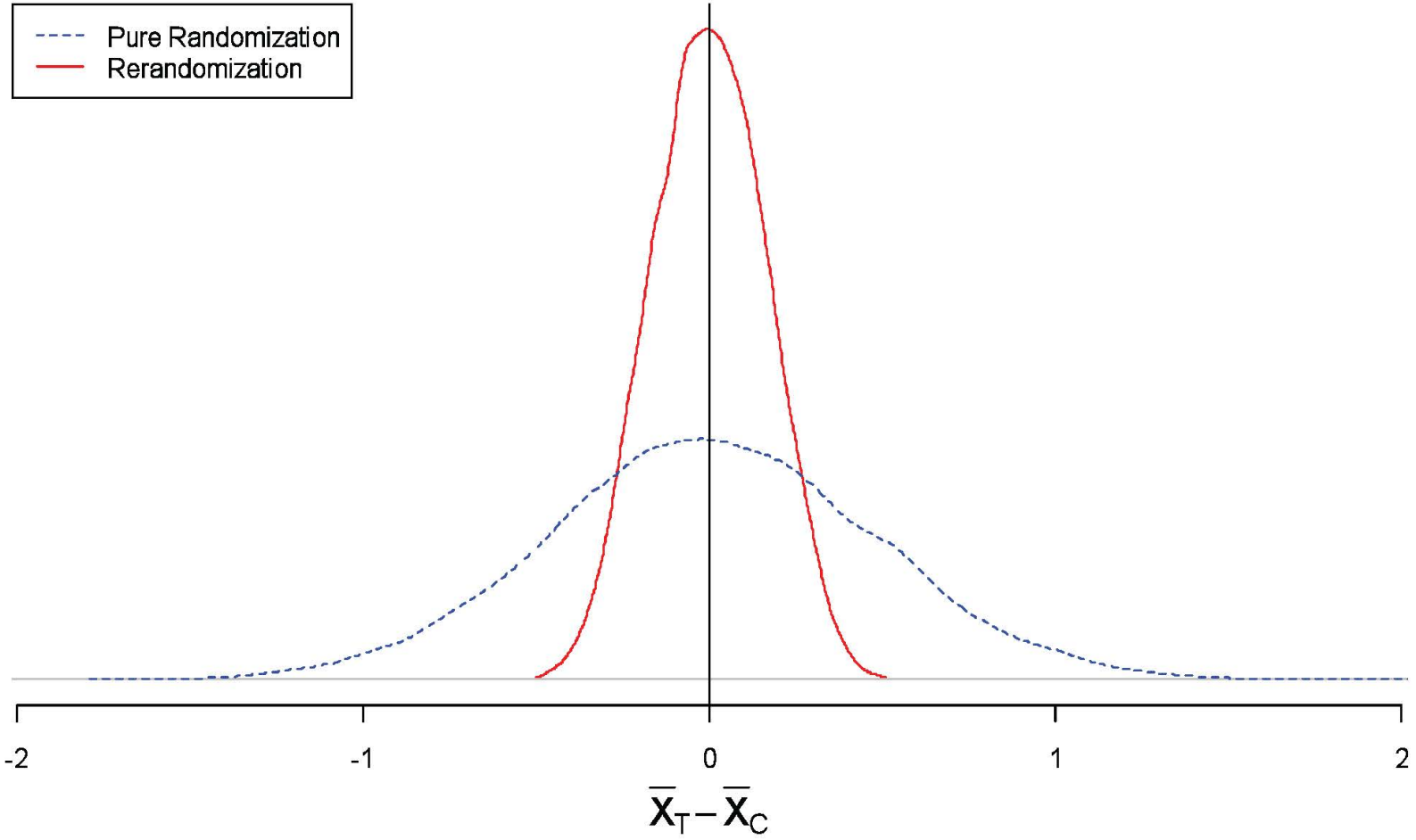


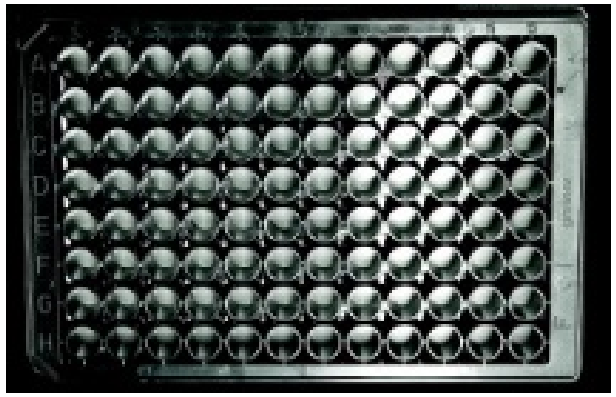
Figure courtesy: Kari Lock Morgan and Donald B. Rubin

The Stem-cell Experiment

- Type 1 Diabetes and Regenerative Medicine
- Converting stem cells to insulin-generating beta cells using chemical modulators

Experimental unit

- A well containing about 20,000 stem cells. Arranged in 8x12 array in plates (96 wells per plate).
- Each well will receive treatment or control (apply modulator or not)
- Response: Percentage of cells that are converted to beta cells (after a specified time following the experiment)



Assigning a single modulator to units

PROBLEM: Assign treatment to <u>two</u> 3x4 arrays (plates) of wells (identify 12 wells that receive treatment and 12 that do not)

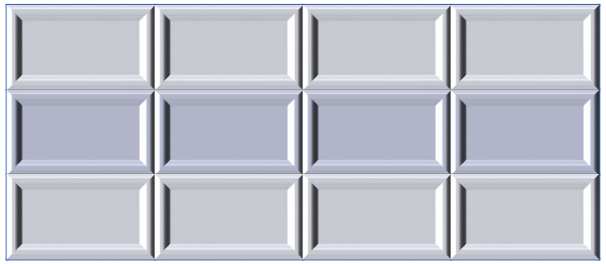


Plate 1

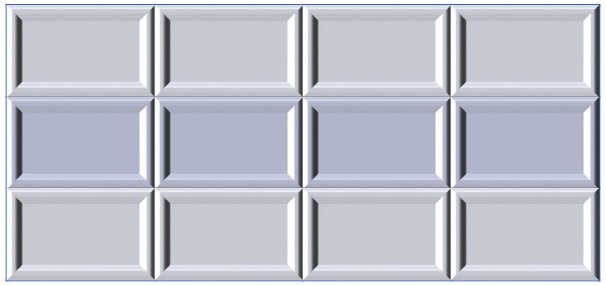


Plate 2

Assigning a single modulator to units

PROBLEM: Assign treatment to <u>two</u> 3x4 arrays (plates) of wells (identify 12 wells that receive treatment and 12 that do not)

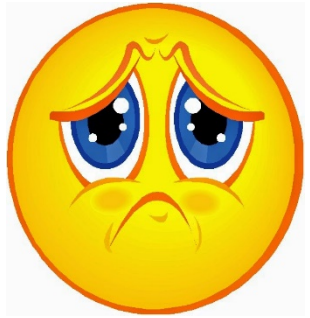
1	1	1	1
1	1	1	1
1	1	1	1

Plate 1

0	0	0	0
0	0	0	0
0	0	0	0

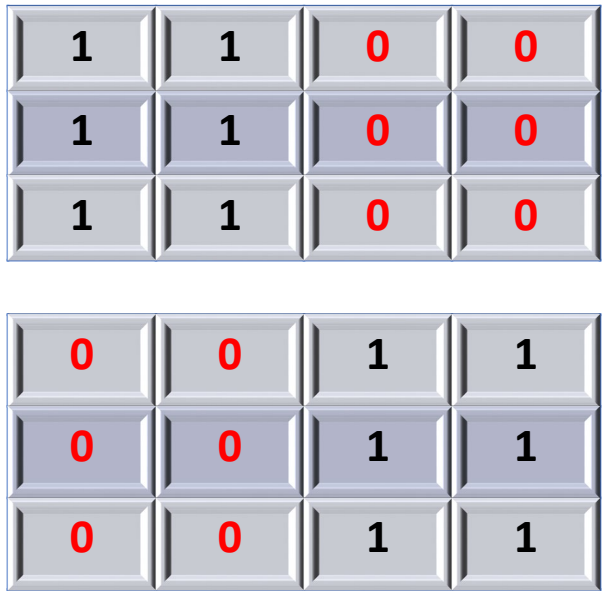
Plate 2

Treatment effect likely to be confounded with factors affecting plate-to-plate variation!



Avoiding bad assignments: Blocking

Within each plate, assign six units to treatment

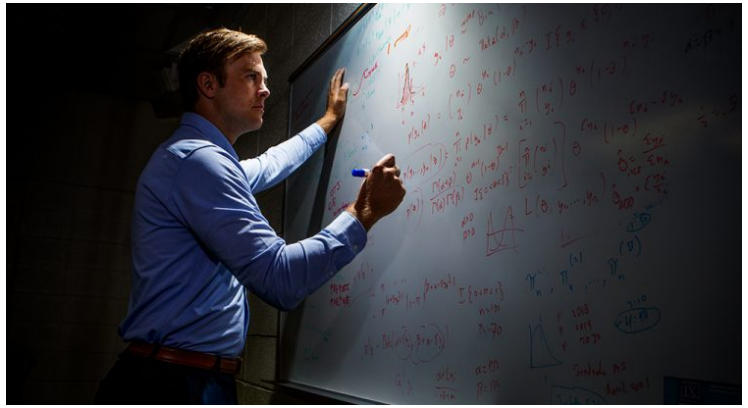


GOOD ASSIGNMENT ?

STATISTICIAN'S SUGGESTION: RANDOMIZE!

A hypothetical conversation (usually not mentioned)

1	1	0	0
1	1	0	0
1	1	0	0



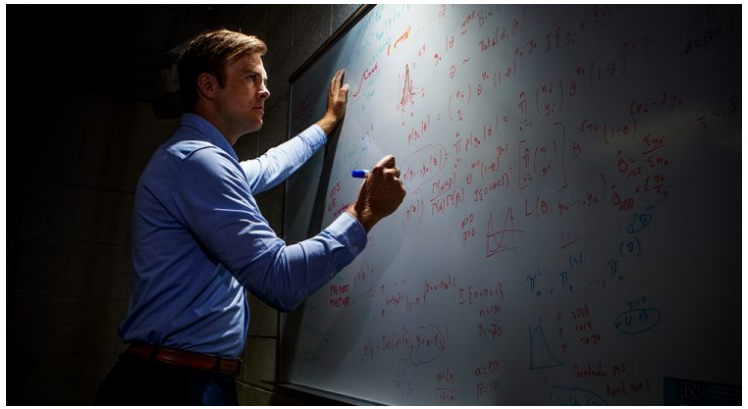
BIOTECHNOLOGIST:
My student came up with this deterministic assignment. Is this OK?

STATISTICIAN:
Oh dear! This is a terrible assignment. I cannot analyze data and draw meaningful conclusions from such an assignment. Could you please ask your student to randomize the assignments?

Images from: https://career.uconn.edu/, http://weusemath.org

A hypothetical conversation (contd.)

1	1	0	0
1	1	0	0
1	1	0	0



BIOTECHNOLOGIST:
After randomization, my student came up with this assignment that seems to be the same as before! I suppose we should scrap it and generate another assignment?

STATISTICIAN:
Oh no! That would be cheating! Since your student obtained this assignment via randomization, it is OK now.

Images from: https://career.uconn.edu/, http://weusemath.org

The million dollar question ….

How can the same assignment be acceptable if it is generated by randomization and unacceptable if not?

Define an acceptable randomization apriori!

Pre-define a measure of balance and acceptable randomization

- Average distance between pairs of points assigned to control (0)
- Hamming distance, e.g. distance between [1,1] and [2,1] is 1, distance between [1,3] and [2,4] is 2.
- A randomization is acceptable if the average Hamming distance between units assigned to control exceeds 1.5

1	1	0	0
1	1	0	0
1	1	0	0

Avg Hamming distance: 1.4

0	1	0	0
0	1	1	0
1	0	1	1

Avg Hamming distance: 1.6

The randomization test: generate all "acceptable" randomizations

1	1	1	0
0	1	1	0
1	0	0	0

0.60	0.36	0.13	0.48
0.46	0.48	0.36	0.33
0.07	0.55	0.82	0.75

t^{rep}=-2.92

1	0	1	0
1	0	1	0
1	0	0	1

0.60	0.36	0.13	0.48
0.46	0.48	0.36	0.33
0.07	0.55	0.82	0.75

t^{rep}=-1.18

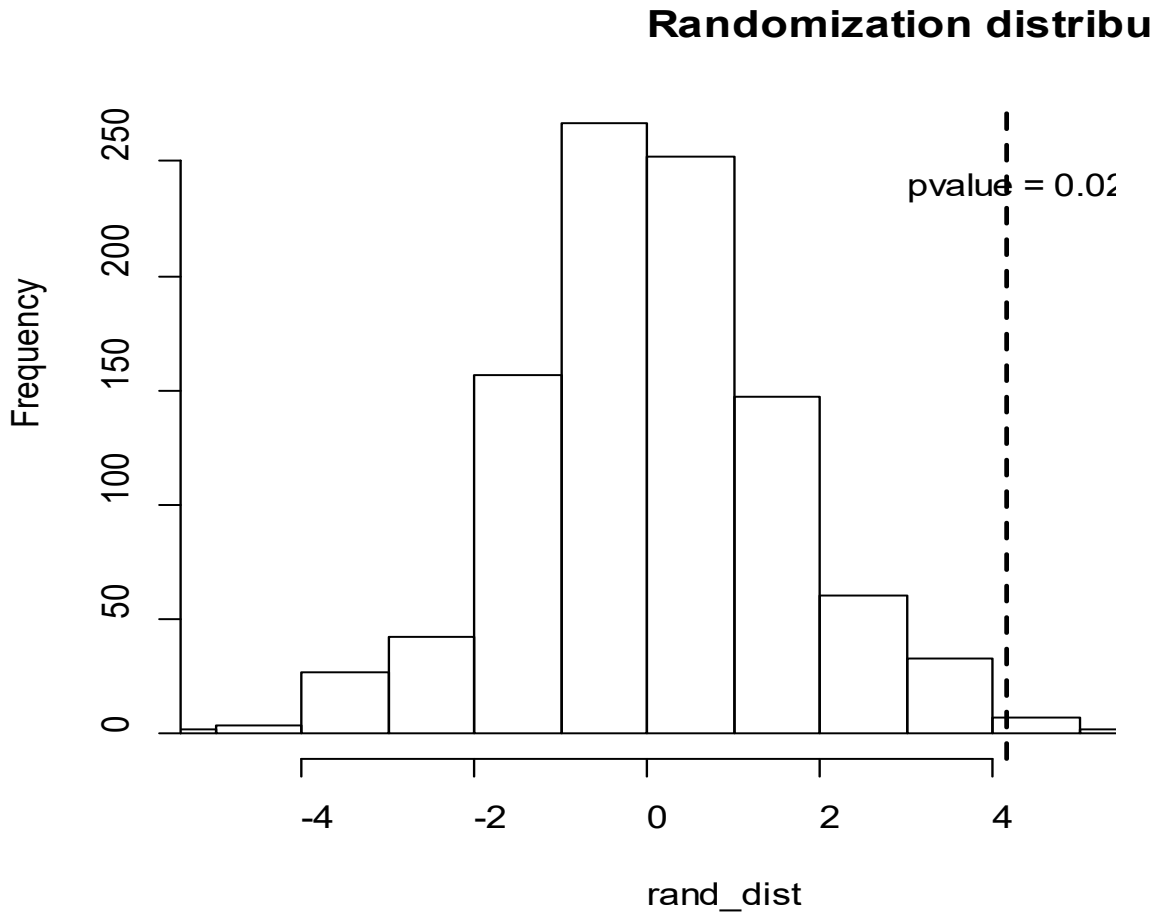
0	1	0	0
0	1	1	0
1	0	1	1

0.60	0.36	0.13	0.48
0.46	0.48	0.36	0.33
0.07	0.55	0.82	0.75

t^{rep}=-0.51

Data remains same, only repeated assignments are generated using the same mechanism as the one used in the experiment to obtain the data

The randomization test



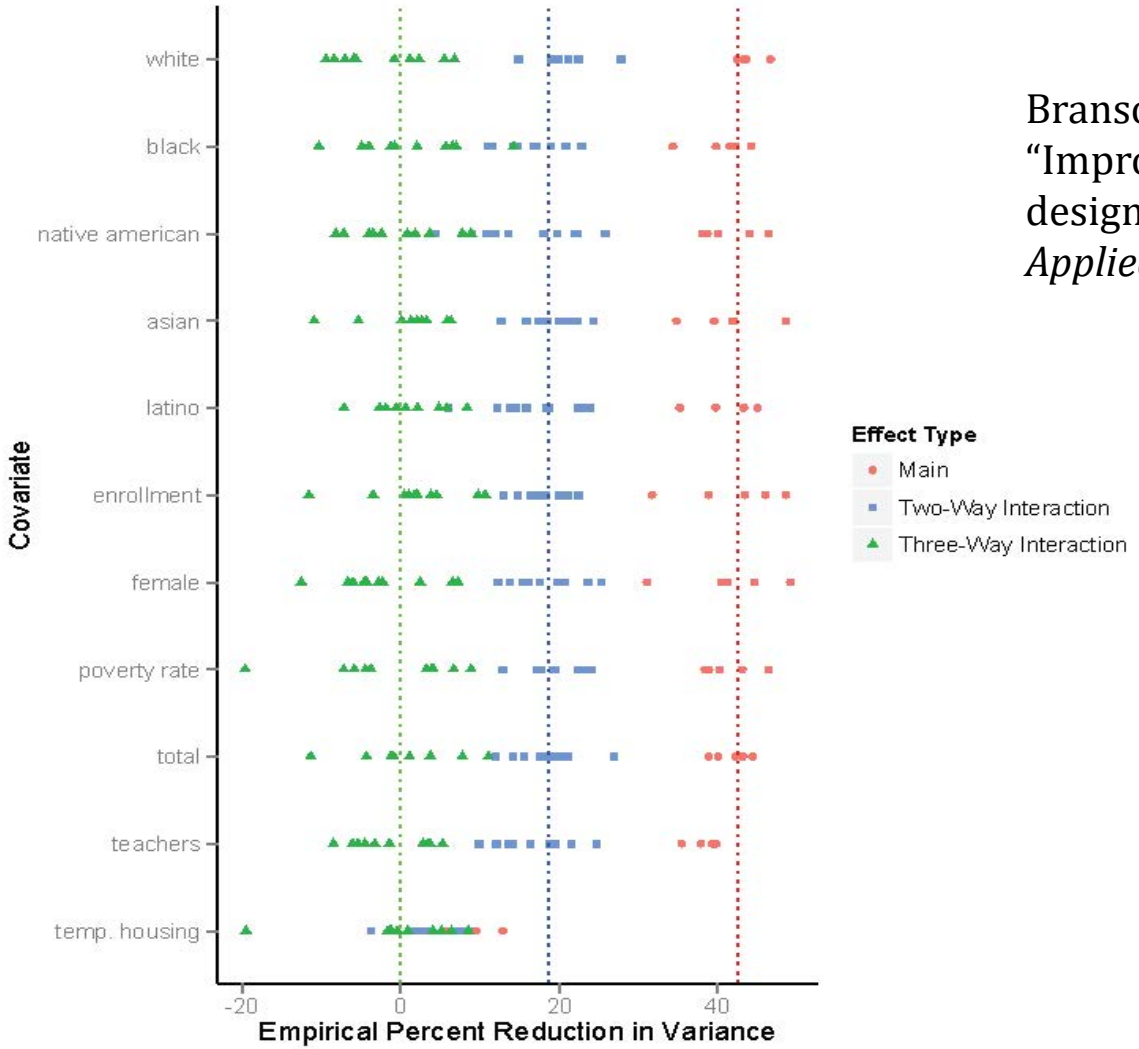
Multi-factor experiments

- 224 New York schools
- Five new interventions labelled A-E, e.g.,
  - Quality review (A)
  - School-wide performance bonus scheme for the teachers (B)
- Response: A cumulative score on the annual progress report.
- A 2^5 factorial experiment with five factors each at two levels: 1(treatment), 0 (control).

Acceptable assignments

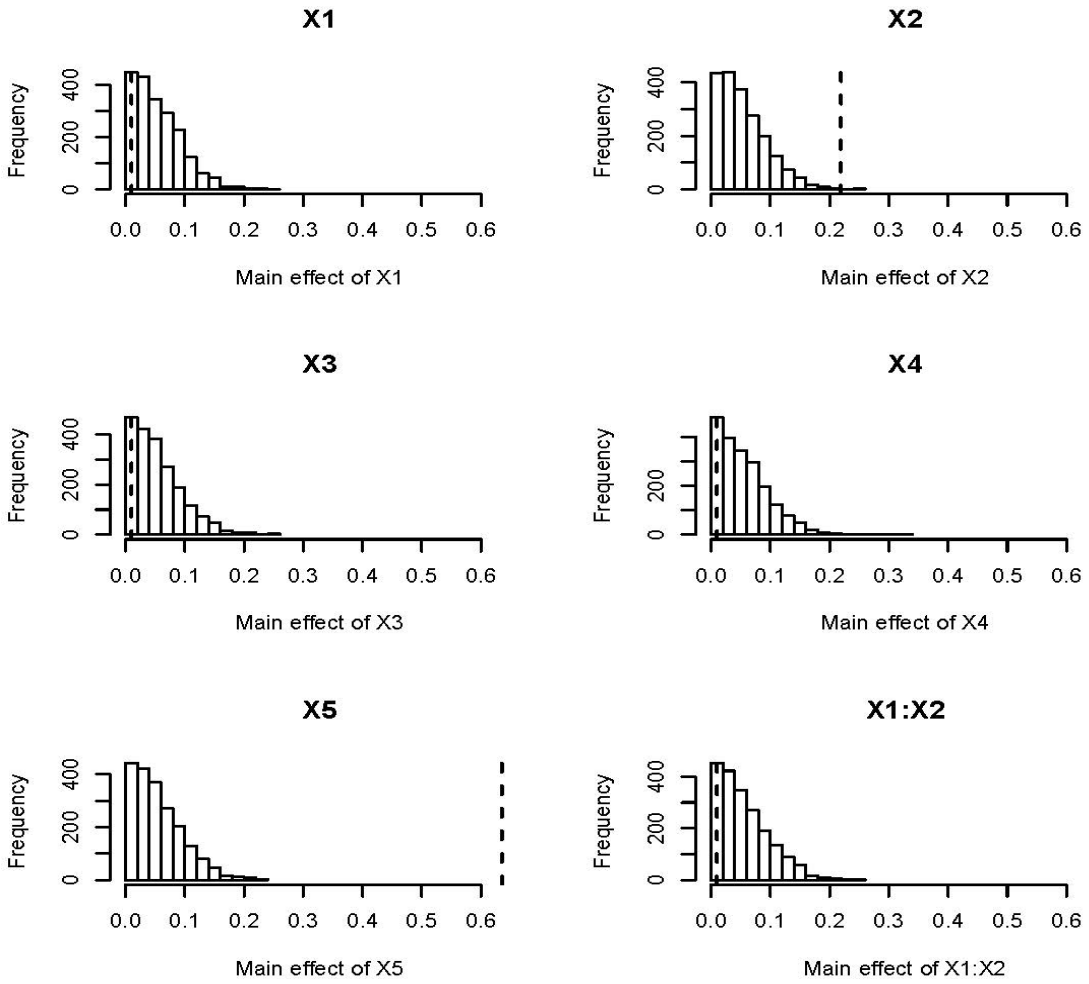
- Completely randomized assignment (CRA) of the 32 treatment combinations to the 224 schools (each treatment to eight schools).
- But need balance over 50 covariates
- Different levels of protection (balance):
  - Maximum protection to five main effects
  - Less protection to two-factor interactions
  - Zero protection to three, four, five-factor interactions

Improving balance through acceptable randomizations



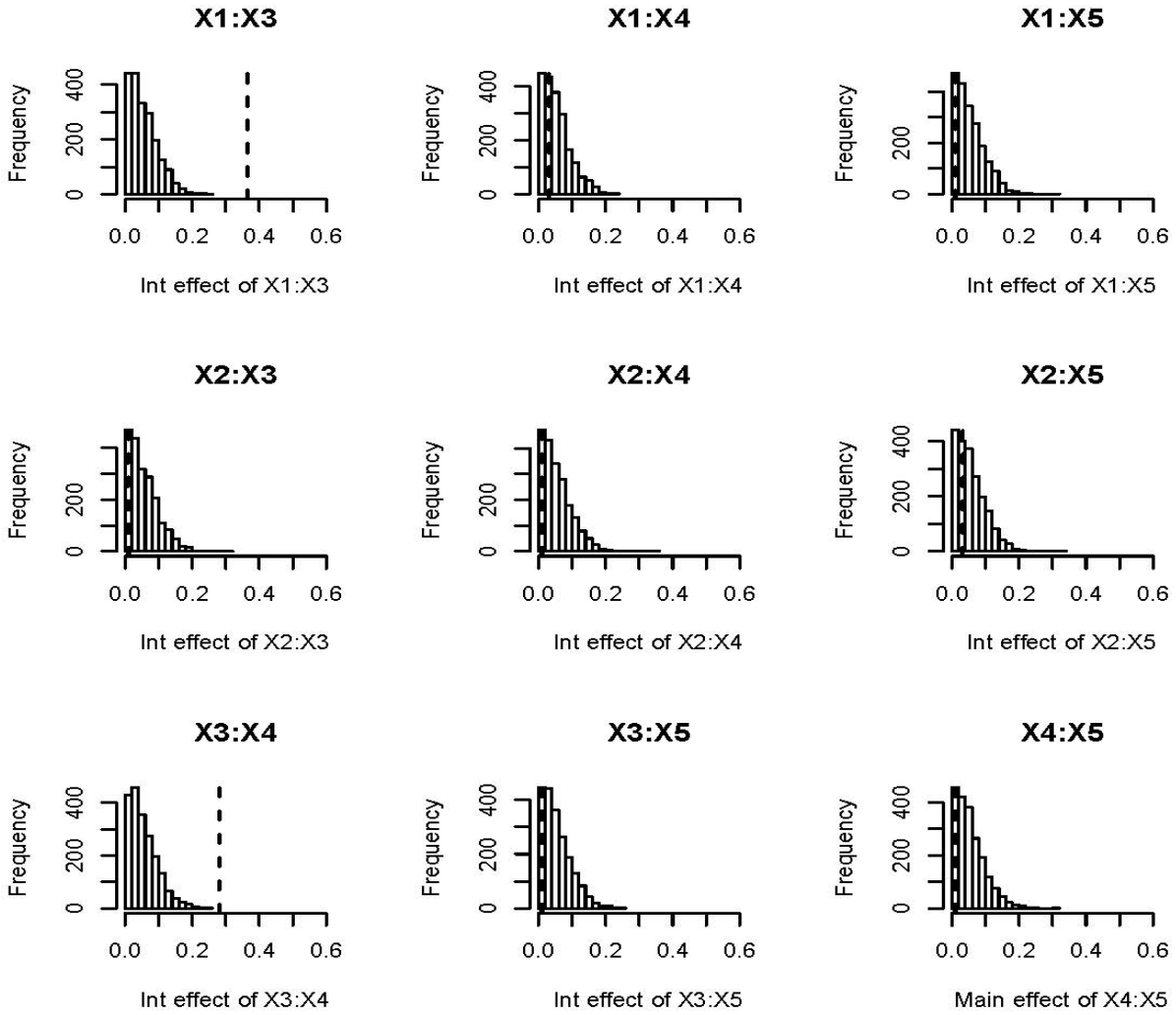
Branson, Z., Dasgupta, T. and Rubin, D.B. (2017) "Improving Covariate Balance in 2^K factorial designs via Re-randomization," *The Annals of Applied Statistics*, 10, 1958-1976.

Randomization tests



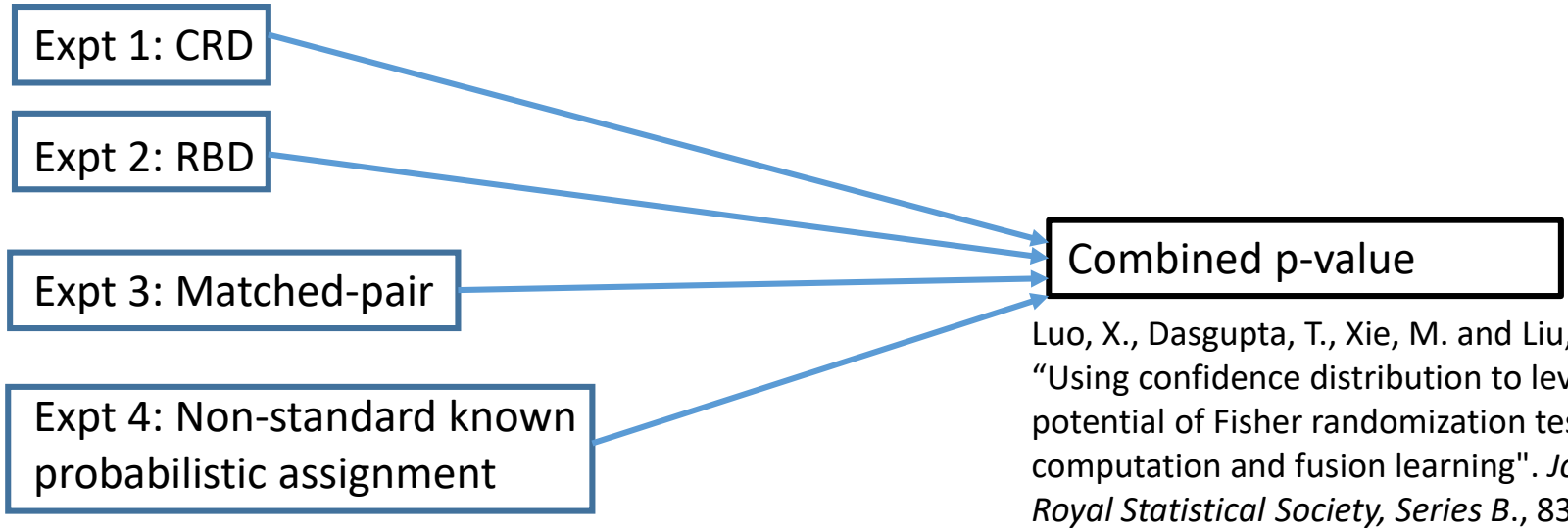
Dasgupta, T., Pillai, N. and Rubin, D.R. (2015), "Causal Inference for 2^K factorial designs by using potential outcomes," *Journal of the Royal Statistical Society, Series B*, 77(4), 727-753.

Randomization tests (contd.)



Randomized experiments in the age of data fusion

Need to “combine results from many experimental and observational studies, each conducted on a different population and under a different set of conditions in order to synthesize an aggregate measure of targeted effect size that is better, in some sense, than any one study in isolation." Pearl (2016), in a recent PNAS issue on causal inference from big data.



Luo, X., Dasgupta, T., Xie, M. and Liu, R. (2021), “Using confidence distribution to leverage the potential of Fisher randomization tests: inference, computation and fusion learning". *Journal of the Royal Statistical Society, Series B*., 83, 777-797.

Why are randomization tests great?

- Intuitive - analyze as you randomize. Easy to teach.
- Flexibility and broad applicability
  - Continuous/binary response
  - ANY test statistic
  - ANY probabilistic assignment mechanism (beyond standard designs)
  - multiple factors
- A “valid” test of the sharp null hypothesis of equal treatment effect for all experimental units
- Possible to invert to obtain confidence intervals
- A Bayesian connection (Rubin 1984) Also see Espinosa, V., Dasgupta, T. and Rubin, D. B. (2016), “A Bayesian perspective on the analysis of unreplicated factorial designs using potential outcomes,” *Technometrics,* 58, 62-73

Some other references

- Ding, P. and Dasgupta, T. (2016), “A Potential Tale of Two by Two Tables from Completely Randomized Experiments”, *Journal of the American Statistical Association (Theory and Methods)*, 111, 157-168.
- Ding, P. and Dasgupta, T. (2018) “A randomization-based perspective of analysis of variance: a test statistic robust to treatment effect heterogeneity,” *Biometrika*, 105 (1), 45-56.
- Luo, X., Dasgupta, T., Xie, M. and Liu, R. (2021), “Using confidence distribution to leverage the potential of Fisher randomization tests: inference, computation and fusion learning". *Journal of the Royal Statistical Society, Series B*., 83, 777-797.