Experimental design and its role in data science

Tirthankar Dasgupta

CS 109 / Stat 121

November 17, 2015

Using data to answer a “causal” question

- 100 people with headaches took an aspirin each, and 90 of them were cured after an hour.

- Does this suggest aspirin relieves headache within an hour?

No: ppl would’ve gotten better w/o aspirin?
100 ppl = small sample size.

Bad question: what is scope of my inference?

Image: www.123rf.com

Does “Big Data” help?

- Ten million people with headaches took an aspirin each, and nine million of them were cured after an hour.

- Does this suggest aspirin relieves headache within an hour?

Image: www.123rf.com

The “missing” data

- 100 people with headaches took an aspirin each, and 90 of them were cured after an hour.

- What would have happened to these people if they did nothing (or simply drank plenty of water and/or simply relaxed for an hour)?

Adding a “control” group

Treatment	Effect after one hour	
	Cured	Not cured
Took aspirin (Treatment)	90	10
Did nothing (Control)	80	20

Conclusions?

Are control and treatment groups comparable? (ie, random assignment)
Do we have selection bias?

Stronger evidence?

Treatment	**Effect after one hour**	
	Cured	Not cured
Took aspirin (Treatment)	90	10
Did nothing (Control)	50	50

Even stronger evidence?

Treatment	**Effect after one hour**	
	Cured	Not cured
Took aspirin (Treatment)	90	10
Did nothing (Control)	10	90

But what if …..

- 100 individuals exposed to treatment are:

- 100 individuals exposed to control are:

Images: www.vectorstock.com, www.dreamstime.com

“Designing” the study

- Define your objective (does aspirin cure headache in an hour?)
  - Formulate question in “data science” language
  - Identify scope of inference (“whose headache?”)
- Need a treatment group and a control group.
  - Assignment mechanism

need to resemble each other (mimic outcome of same person in two states unobservable).

- These groups should be “identical” (how to define identical and how to achieve?)
  - Big data challenge: large number of covariates associated with each experimental unit.

Analyzing the outcomes

- *Related to* and *consistent with* the design.
- How to calculate the "strength" of your conclusion?

hypothesis testing, p-values (significance).

Treatment	Effect after one hour	
	Cured	Not cured
Took aspirin (Treatment)	70	30
Did nothing (Control)	50	50

Treatment	Effect after one hour	
	Cured	Not cured
Took aspirin (Treatment)	90	10
Did nothing (Control)	10	90

Modern-day experiments

- Education new interventions: causal impact on school performance?
- Marketing focus groups: determining most popular of test products.
- Stem Cell insulin produced by stem cells as diabetes treatment: what is optimal set-up of stem cells?
- Nanotechnology
- Law
- Internet: A/B testing

  new ad: ie, website creates new intervention and tries to judge reaction

  A/B testing similar to control/treatment framework.

How it all started

- Agricultural experiments at the Rothamstead experimental station, U.K.
- Sir R. A. Fisher hired in 1919, first edition of Design of Experiments" published in 1935.

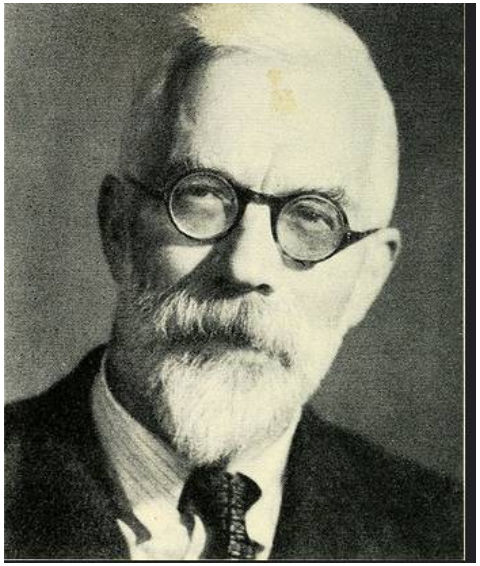
Aerial view of Rothamstead in 2013 (www.bbsrc.ac.uk)

Treatment effect = E[Y_i|X = 1] - E[Y_i|X = 0] —> difference in outcome on SAME person
This is unobservable.

Potential Outcomes

A potential outcome for each unit when exposed to each treatment level

Jerzy Neyman: originated the concept (1923) and introduced the first formal notation



R. A. Fisher (1918): If we say “this boy is tall because he has been well fed”, we are suggesting that he might quite probably have been worse fed, and that in this case he would been shorter.

The fertilizer experiment (does a new fertilizer improve yield of tomatoes?)

Plot of land	Fertilizer A (old)	Fertilizer B (new)	Unit-level effect
1	$Y_1(c)$	$Y_1(t)$	$\tau_1 = Y_1(t)-Y_1(c)$
2			
3			
4			
5			
6			
7			
8			
9			
10			
11			

The fertilizer experiment (does a new fertilizer improve yield of tomatoes?)

Plot of land	Fertilizer A (old)	Fertilizer B (new)	Unit-level effect
1	$Y_1(c)$	$Y_1(t)$	$\tau_1 = Y_1(t)-Y_1(c)$
2	$Y_2(c)$	$Y_2(t)$	$\tau_2 = Y_2(t)-Y_2(c)$
3			
4			
5			
6			
7			
8			
9			
10			
11			

The fertilizer experiment (does a new fertilizer improve yield of tomatoes?)

Plot of land	Fertilizer A (old)	Fertilizer B (new)	Unit-level effect
1	$Y_1(c)$	$Y_1(t)$	$\tau_1 = Y_1(t)-Y_1(c)$
2	$Y_2(c)$	$Y_2(t)$	$\tau_2 = Y_2(t)-Y_2(c)$
3	$Y_3(c)$	$Y_3(t)$	$\tau_3 = Y_3(t)-Y_3(c)$
4	$Y_4(c)$	$Y_4(t)$	$\tau_4 = Y_4(t)-Y_4(c)$
5	$Y_5(c)$	$Y_5(t)$	$\tau_5 = Y_5(t)-Y_5(c)$
6	$Y_6(c)$	$Y_6(t)$	$\tau_6 = Y_6(t)-Y_6(c)$
7	$Y_7(c)$	$Y_7(t)$	$\tau_7 = Y_7(t)-Y_7(c)$
8	$Y_8(c)$	$Y_8(t)$	$\tau_8 = Y_8(t)-Y_8(c)$
9	$Y_9(c)$	$Y_9(t)$	$\tau_9 = Y_9(t)-Y_9(c)$
10	$Y_{10}(c)$	$Y_{10}(t)$	$\tau_{10} = Y_{10}(t)-Y_{10}(c)$
11	$Y_{11}(c)$	$Y_{11}(t)$	$\tau_{11} = Y_{11}(t)-Y_{11}(c)$

The fertilizer experiment (does a new fertilizer improve yield of tomatoes?)

Plot of land	Fertilizer A (old)	Fertilizer B (new)	Unit-level effect
1	$Y_1(c)$	$Y_1(t)$	$\tau_1 = Y_1(t)-Y_1(c)$
2	$Y_2(c)$	$Y_2(t)$	$\tau_2 = Y_2(t)-Y_2(c)$
3	$Y_3(c)$	$Y_3(t)$	$\tau_3 = Y_3(t)-Y_3(c)$
4	$Y_4(c)$	$Y_4(t)$	$\tau_4 = Y_4(t)-Y_4(c)$
5	$Y_5(c)$	$Y_5(t)$	$\tau_5 = Y_5(t)-Y_5(c)$
6	$Y_6(c)$	$Y_6(t)$	$\tau_6 = Y_6(t)-Y_6(c)$
7	$Y_7(c)$	$Y_7(t)$	$\tau_7 = Y_7(t)-Y_7(c)$
8	$Y_8(c)$	$Y_8(t)$	$\tau_8 = Y_8(t)-Y_8(c)$
9	$Y_9(c)$	$Y_9(t)$	$\tau_9 = Y_9(t)-Y_9(c)$
10	$Y_{10}(c)$	$Y_{10}(t)$	$\tau_{10} = Y_{10}(t)-Y_{10}(c)$
11	$Y_{11}(c)$	$Y_{11}(t)$	$\tau_{11} = Y_{11}(t)-Y_{11}(c)$
Average	$\bar{Y}(c)$	$\bar{Y}(t)$	$\tau = \bar{Y}(t) - \bar{Y}(c) = \sum \tau_i / 11$

mean: linear fct, computationally easy.

Median is harder (need cdf = 0.5, pdf hard to integrate)

The “assignment mechanism” and observed outcomes

Plot of land	Fertilizer A (old)	Fertilizer B (new)	Assignment (W)
1	29.2	?	0
2			
3			
4			
5			
6			
7			
8			
9			
10			
11			
Average			

The assignment mechanism and observed outcomes (contd.)

Plot of land	Fertilizer A (old)	Fertilizer B (new)	Assignment (W)
1	29.2	?	0
2	11.4	?	0
3			
4			
5			
6			
7			
8			
9			
10			
11			
Average			

Random assignment of treatment: W. Perhaps constrain to balance?

The assignment mechanism and observed outcomes (contd.)

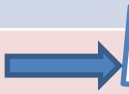
Plot of land	Fertilizer A (old)	Fertilizer B (new)	Assignment (W)
1	29.2	?	0
2	11.4	?	0
3	?	26.6	1
4	?	23.7	1
5	25.3	?	0
6	?	28.5	1
7	?	14.2	1
8	?	17.9	1
9	16.5	?	0
10	21.1	?	0
11	?	24.3	1
Average	20.70	22.53	Diff = 1.83

Fisher's "sharp" null hypothesis

- No effect of fertilizer on ANY plot
- How to assess?
- Stochastic proof by contradiction!
  - Calculate observed value of test statistic
  - Assuming hypothesis to be true, impute missing potential outcomes
  - Generate distribution of test statistic using repeated assignments under same mechanism
  - Determine if observed value is "unusual"

Step-1: Calculate observed value of test statistic

Plot of land	Fertilizer A (old)	Fertilizer B (new)	Assignment (W)
1	29.2	?	0
2	11.4	?	0
3	?	26.6	1
4	?	23.7	1
5	25.3	?	0
6	?	28.5	1
7	?	14.2	1
8	?	17.9	1
9	16.5	?	0
10	21.1	?	0
11	?	24.3	1
Average	20.70	22.53	



$|22.53-20.70| = 1.83$

Step-2: Impute missing potential outcomes under the null hypothesis

Plot of land	Fertilizer A (old)	Fertilizer B (new)	Assignment (W)
1	29.2	29.2	
2	11.4	11.4	
3	26.6	26.6	
4	23.7	23.7	
5	25.3	25.3	
6	28.5	28.5	
7	14.2	14.2	
8	17.9	17.9	
9	16.5	16.5	
10	21.1	21.1	
11	24.3	24.3	
Average			

Step-3: Generate new assignment, new observed outcomes, new value of test statistic

Plot of land	Fertilizer A (old)	Fertilizer B (new)	New Assignment
1	?	29.2	1
2	11.4	?	0
3	?	26.6	1
4	23.7	?	0
5	?	25.3	1
6	?	28.5	1
7	14.2	?	0
8	17.9	?	0
9	?	16.5	1
10	21.1	?	0
11	?	24.3	1
Average	**17.66**	**25.07**	T_{new} **= 7.41**

Step-3 (contd.): Generate new assignment, new observed outcomes, new value of test statistic

Plot of land	Fertilizer A (old)	Fertilizer B (new)	New Assignment
1	29.2	?	0
2	11.4	?	0
3	26.6	?	0
4	23.7	?	0
5	25.3	?	0
6	?	28.5	1
7	?	14.2	1
8	?	17.9	1
9	?	16.5	1
10	?	21.1	1
11	?	24.3	1
Average	**23.24**	**20.42**	T_{new}=2.82

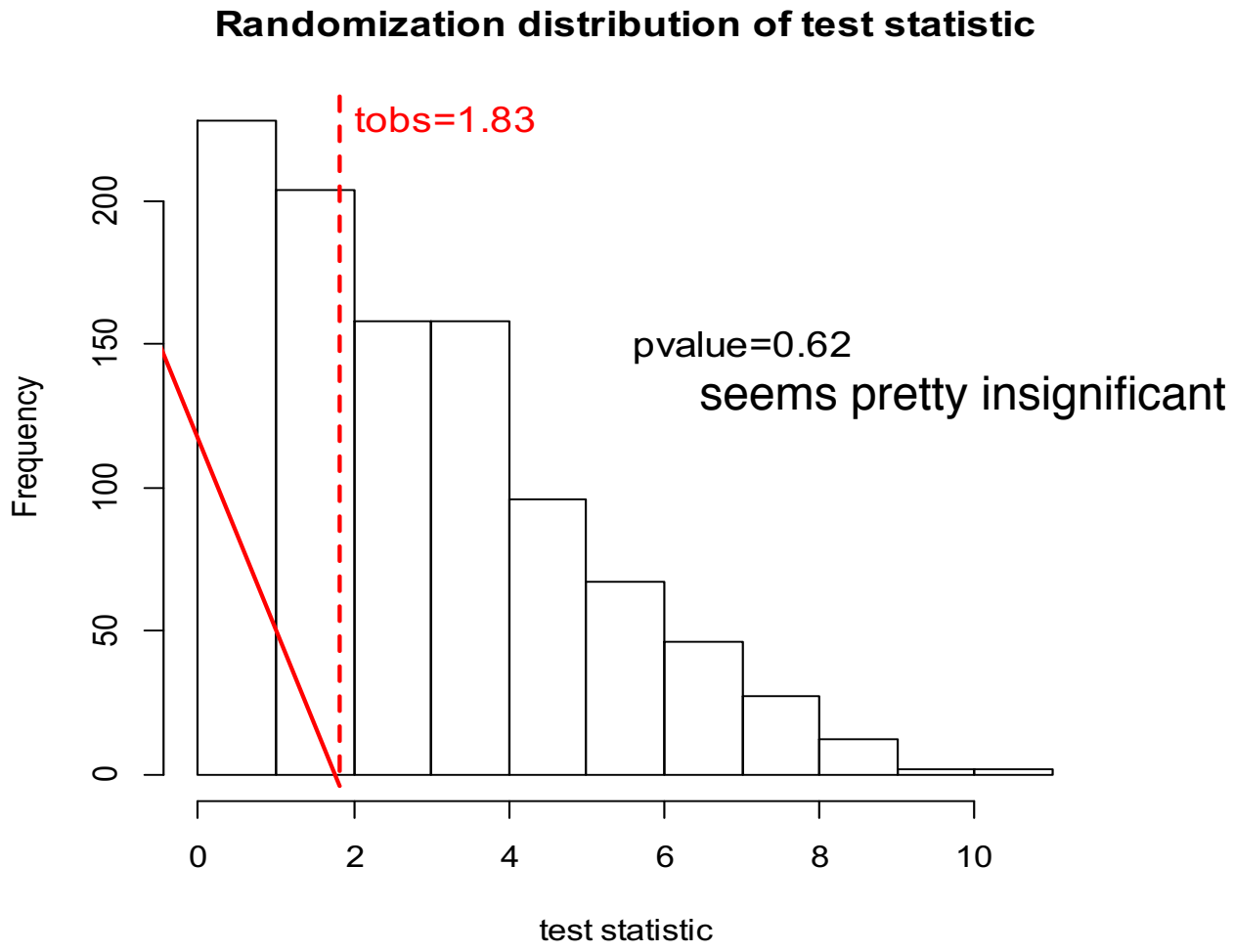
Step-3 (contd.): Generate new assignment, new observed outcomes, new value of test statistic

Plot of land	Fertilizer A (old)	Fertilizer B (new)	New Assignment
1	?	29.2	1
2	?	11.4	1
3	?	26.6	1
4	?	23.7	1
5	?	25.3	1
6	?	28.5	1
7	14.2	?	0
8	17.9	?	0
9	16.5	?	0
10	21.1	?	0
11	24.3	?	0
Average	**18.80**	**24.12**	T_{new}=5.32

How many possible assignments (and hence total values of test statistic)?

$$\frac{11!}{5!6!} = 462$$

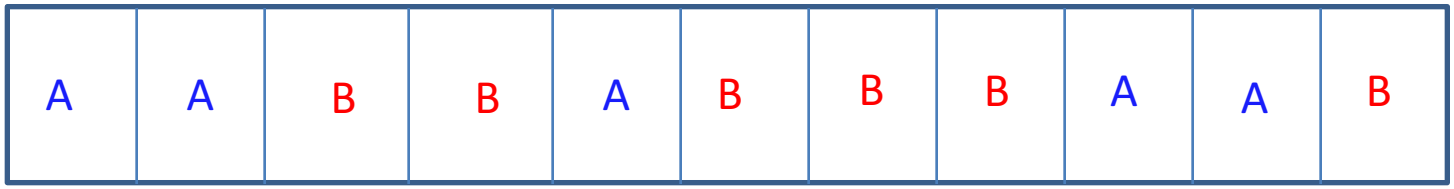
Step-4: Is the observed value of the test statistic “unusual”?



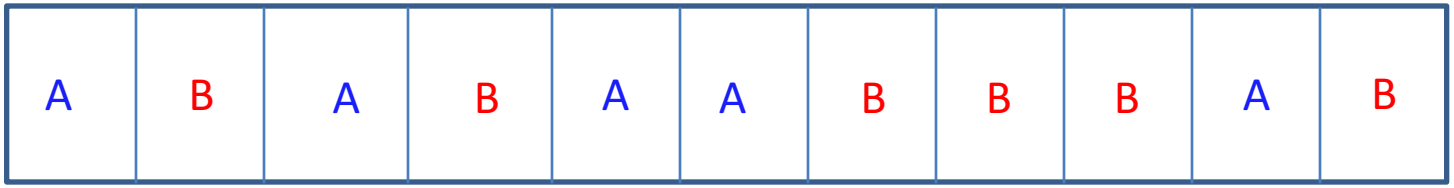
The role of randomization

Which plots receive A and which receive B?

Observed assignment leading to t_{obs}=1.83



But you could have observed this instead:



… and any of the 462 assignments with equal probability

Role of randomization: "Unbiasedness in expectation"

Assignment No.	1	2	3	4	5	6	7	8	9	10	11	Difference of means
1	A	A	A	A	A	B	B	B	B	B	B	-2.82
…	A	A	B	B	A	B	B	B	A	A	B	1.83
462	B	B	B	B	B	B	A	A	A	A	A	5.32

What is the expected value of the statistic over all possible randomizations?

(-2.82+…+1.83+…+5.31)/462 =?

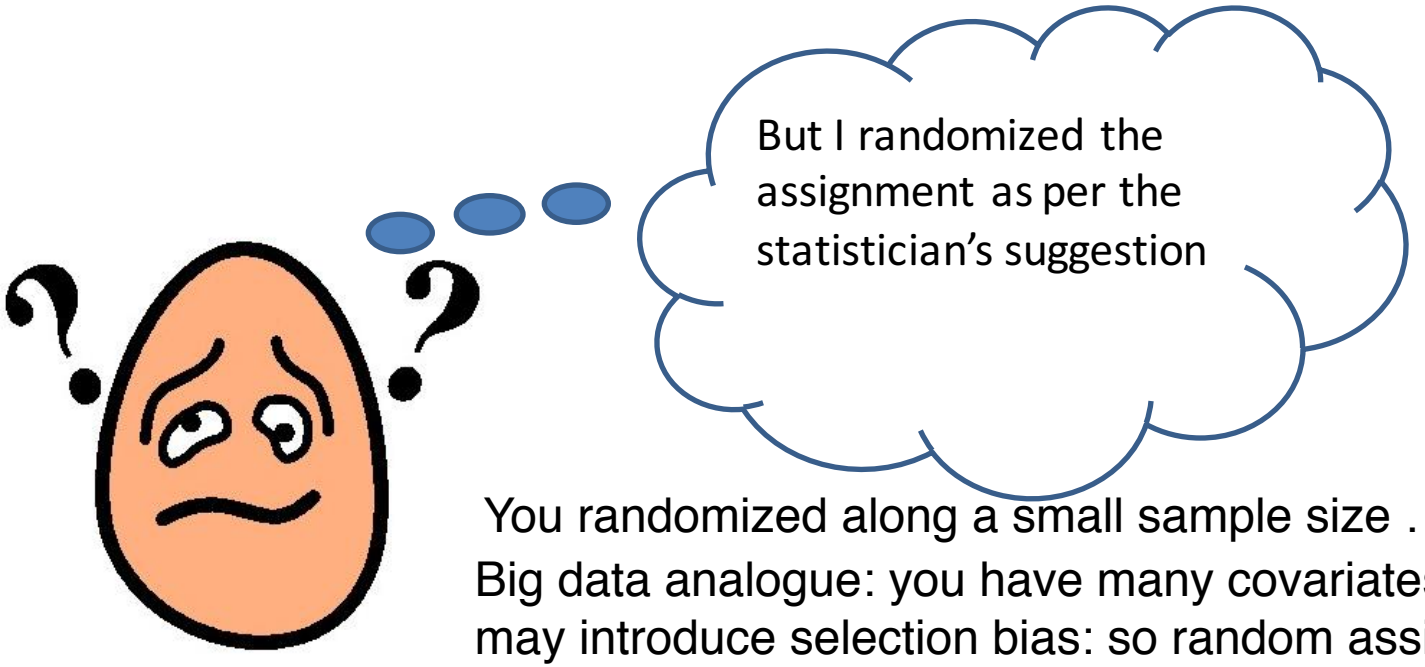
you get the treatment effect: the true difference.

on avg you get the right answer: ONLY protected in expectation.

RANDOMIZATION SAFEGUARDS THE EXPERIMENT AGAINST OBSERVED AND UNOBSERVED COVARIATES IN EXPECTATION (ON AN AVERAGE)

But what if you come up with …

A	A	A	A	A	B	B	B	B	B	B



You randomized along a small sample size …

Big data analogue: you have many covariates which may introduce selection bias: so random assignment is more likely to fail

Blocking: A strategy to protect assignment from "bad" randomization:

FISHER:

- "Block what you can, randomize what you cannot"
- "Analyze as you randomize"

How would our analysis have changed?

The Diet Experiment

- Effect of improved diet A (treatment) versus standard diet B (control).
- Twenty animals available.
- Differ with respect to age, sex and other characteristics.

Matched-pair (blocked) experiment

Extreme blocking (as much blocking as possible).

- Scientist forms 10 pairs of animals.
- Animals in the same pair are “identical”.
- Each animal within each pair gets either diet A or diet B; allocation decided by flip of a coin.

Potential outcomes

PAIR (BLOCK)	Potential outcome for animal 1		Potential outcome for animal 2	
	Diet A	Diet B	Diet A	Diet B
1	$Y_{1,1}(A)$	$Y_{1,1}(B)$	$Y_{1,2}(A)$	$Y_{1,2}(B)$
2	…	…	…	…
3	…	…	…	…
4	…	…	…	…
5	…	…	…	…
6	…	…	…	…
7	…	…	…	…
8	…	…	…	…
9	…	…	…	…
10	$Y_{10,1}(A)$	$Y_{10,1}(B)$	$Y_{10,2}(A)$	$Y_{10,2}(B)$

Observed outcomes

PAIR	Potential outcome for animal 1		Potential outcome for animal 2	
	Diet A	Diet B	Diet A	Diet B
1	13.2	?	?	14.0
2	?	8.8	8.2	?
3	?	11.2	10.9	?
4	14.3	?	?	14.2
5	10.7	?	?	11.8
6	6.6	?	?	6.4
7	?	9.8	9.5	?
8	10.8	?	?	11.3
9	?	9.3	8.8	?
10	?	13.6	13.3	?

animals in pair get opposite treatment: maintain balance in assignment.

Observed value of test statistic

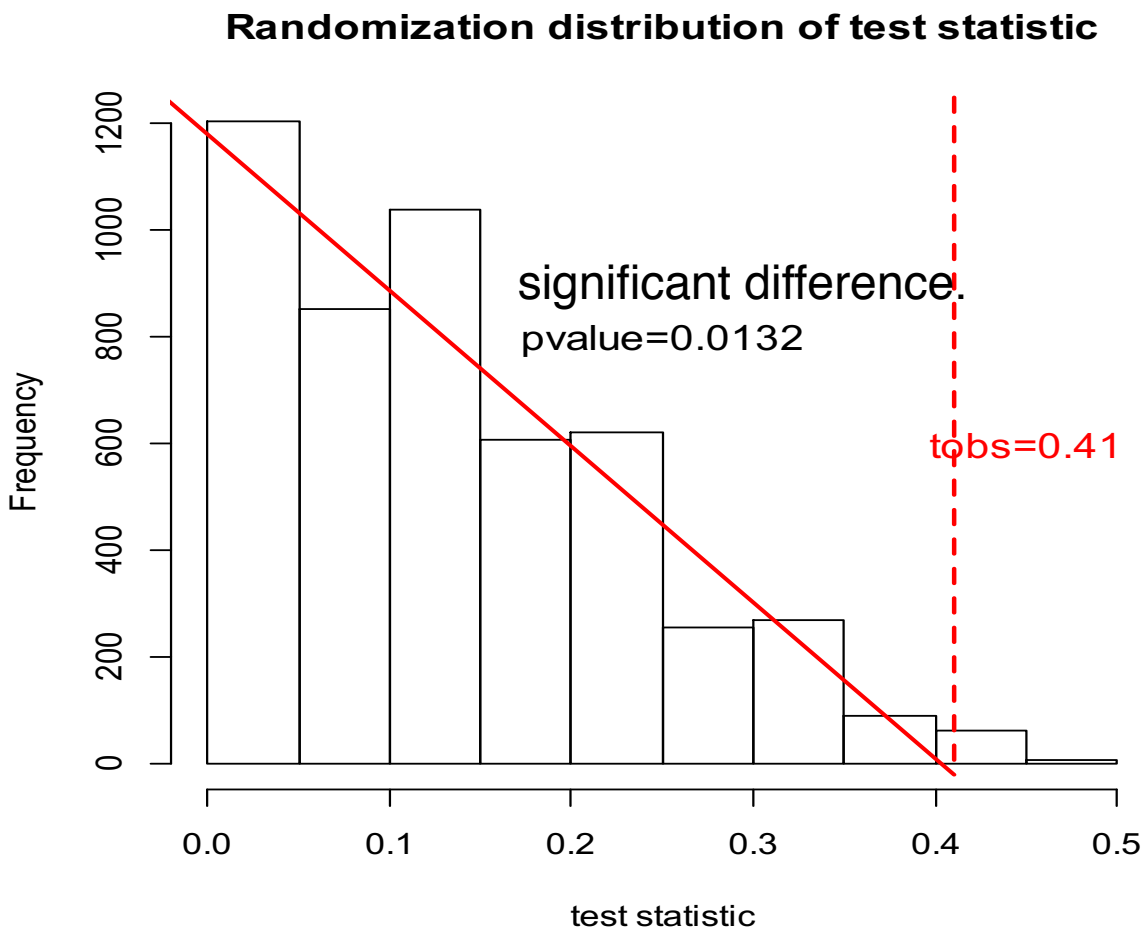
PAIR	Potential outcome for animal 1		Potential outcome for animal 2		Diff (d) [Red – blue]
	Diet A	Diet B	Diet A	Diet B	
1	13.2	?	?	14.0	0.8
2	?	8.8	8.2	?	0.6
3	?	11.2	10.9	?	0.3
4	14.3	?	?	14.2	-0.1
5	10.7	?	?	11.8	1.1
6	6.6	?	?	6.4	-0.2
7	?	9.8	9.5	?	0.3
8	10.8	?	?	11.3	0.5
9	?	9.3	8.8	?	0.5
10	?	13.6	13.3	?	0.3

$$|\bar{d}| = 0.41$$

Imputed table of potential outcomes under sharp null of no effect

PAIR	Potential outcome for animal 1		Potential outcome for animal 2	
	Diet A	Diet B	Diet A	Diet B
1	13.2	13.2	14.0	14.0
2	8.8	8.8	8.2	8.2
3	11.2	11.2	10.9	10.9
4	14.3	14.3	14.2	14.2
5	10.7	10.7	11.8	11.8
6	6.6	6.6	6.4	6.4
7	9.8	9.8	9.5	9.5
8	10.8	10.8	11.3	11.3
9	9.3	9.3	8.8	8.8
10	13.6	13.6	13.3	13.3

Distribution of the test statistic and the p-value



Three fundamental principles of experimentation (Fisher 1925)

- Randomization
- Replication
- Blocking

How experiments are changing

- Hundreds of covariates associated with each experimental unit
  - e.g., patients in clinical trials
- Multiple treatment factors
  - Thirty chemical modulators in stem-cell experiments
- Complex randomization restrictions
  - Non-compliance
  - Multi-stratum

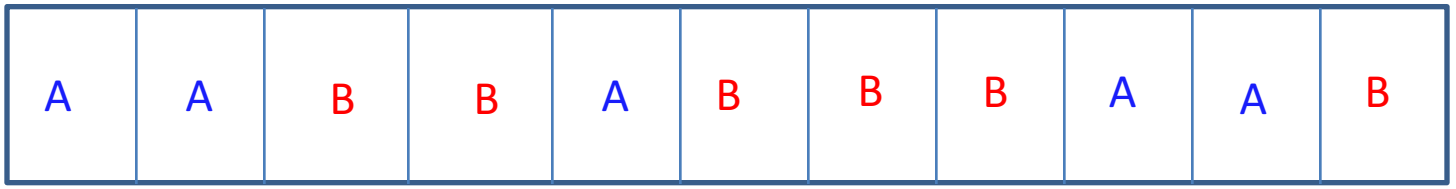
blocking doesn't make sense.
also have effect of covariate interactions:
2 variable interaction, 3 interaction, etc

Designs that balance several covariates over treatment groups

- The more covariates, the more likely at least one covariate will be imbalanced across treatment groups
- Covariate imbalance not limited to “unlucky” randomizations
- Blocking not intuitive
- The solution: **Re**-randomization
  make a decision before the second randomization.

Define a measure of "balance" between treatment and control groups

- Define a measure
- Small values of the measure are acceptable
- Large values of the measure indicate lack of balance and are unacceptable



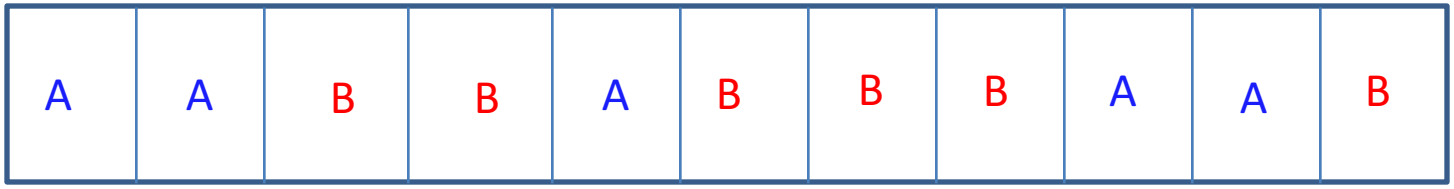
What can be a possible measure of balance?

measure of balance: length of largest sequence?

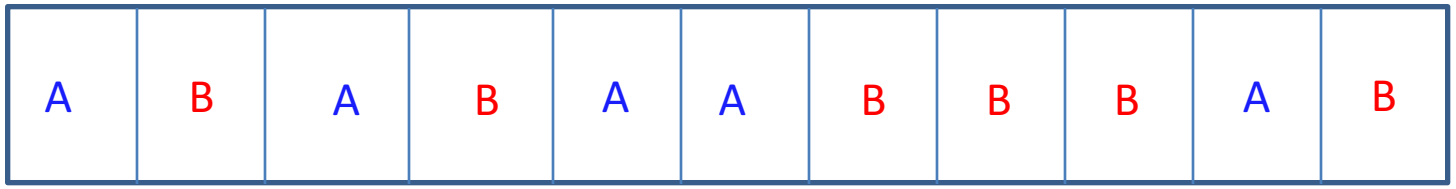
look at positions: avg of positions B and A, find difference

ie, are the A and B spread out evenly (at random - uniform) throughout the data?

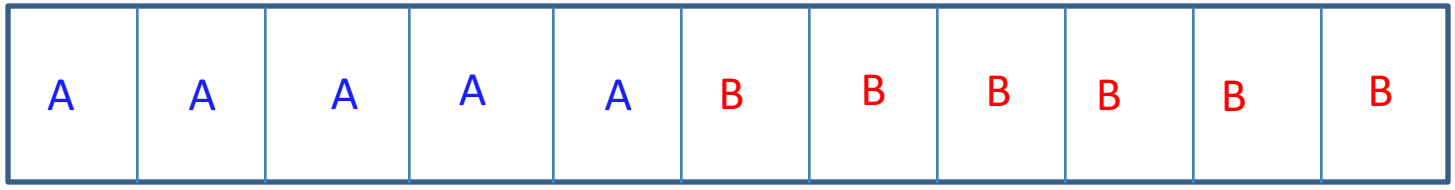
Comparing treatment assignments



(3+4+6+7+8+11)/6-(1+2+5+9+10)/5=1.10



(2+4+7+8+9+11)/6-(1+3+5+6+10)/5=1.83



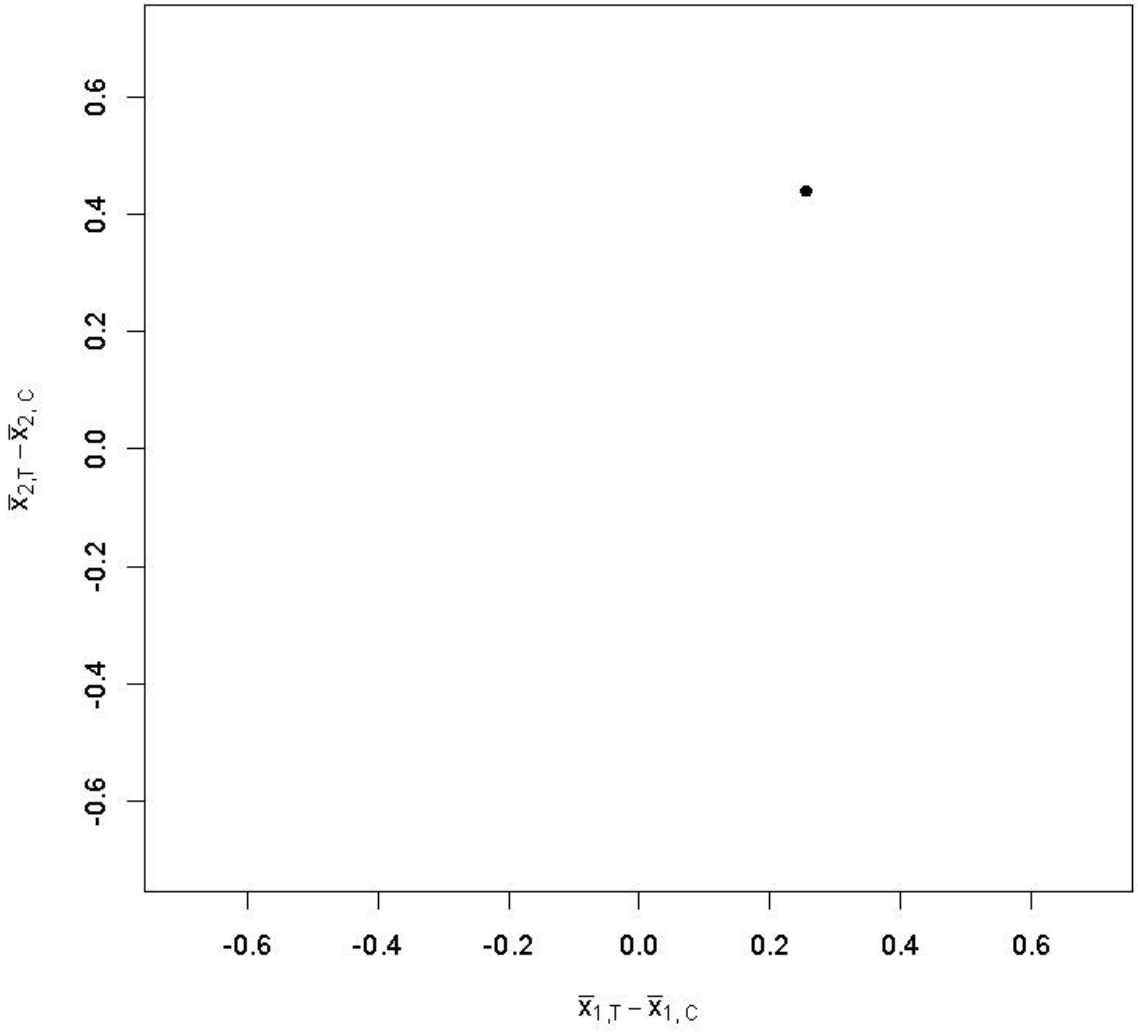
(6+7+8+9+10+11)/6-(1+2+3+4+5)/5=5.5

Little effect on experiment, in expectation.

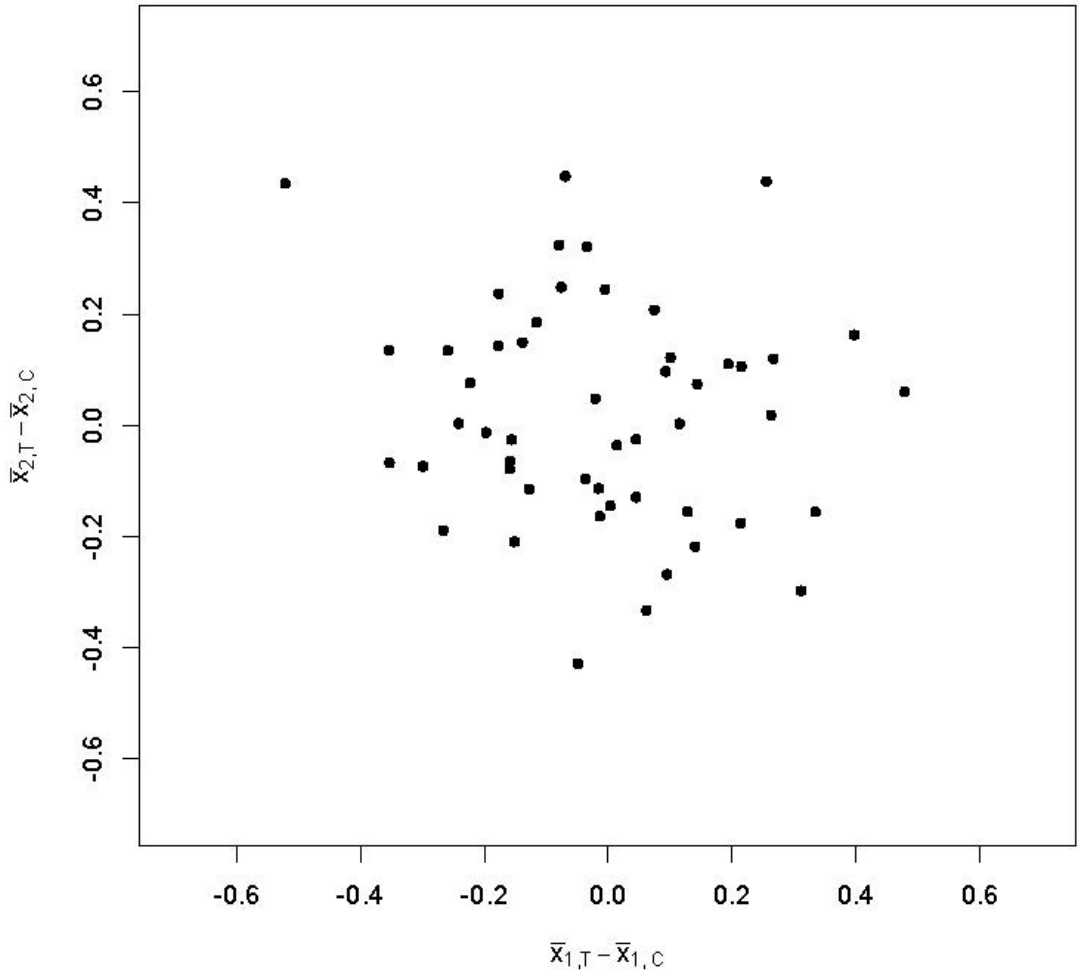
Design and analysis of re-randomized experiments

- Decide acceptable criterion (which randomizations to accept)
- Randomize until the acceptability criterion is met
- Analysis using randomization test:
  - Calculate observed value of test statistic
  - Assuming hypothesis to be true, impute missing potential outcomes
  - Generate distribution of test statistic using repeated assignments under same mechanism (i.e., accepting randomizations that are acceptable)
  - Determine if observed value of "unusual"

Visualization for two continuous covariates

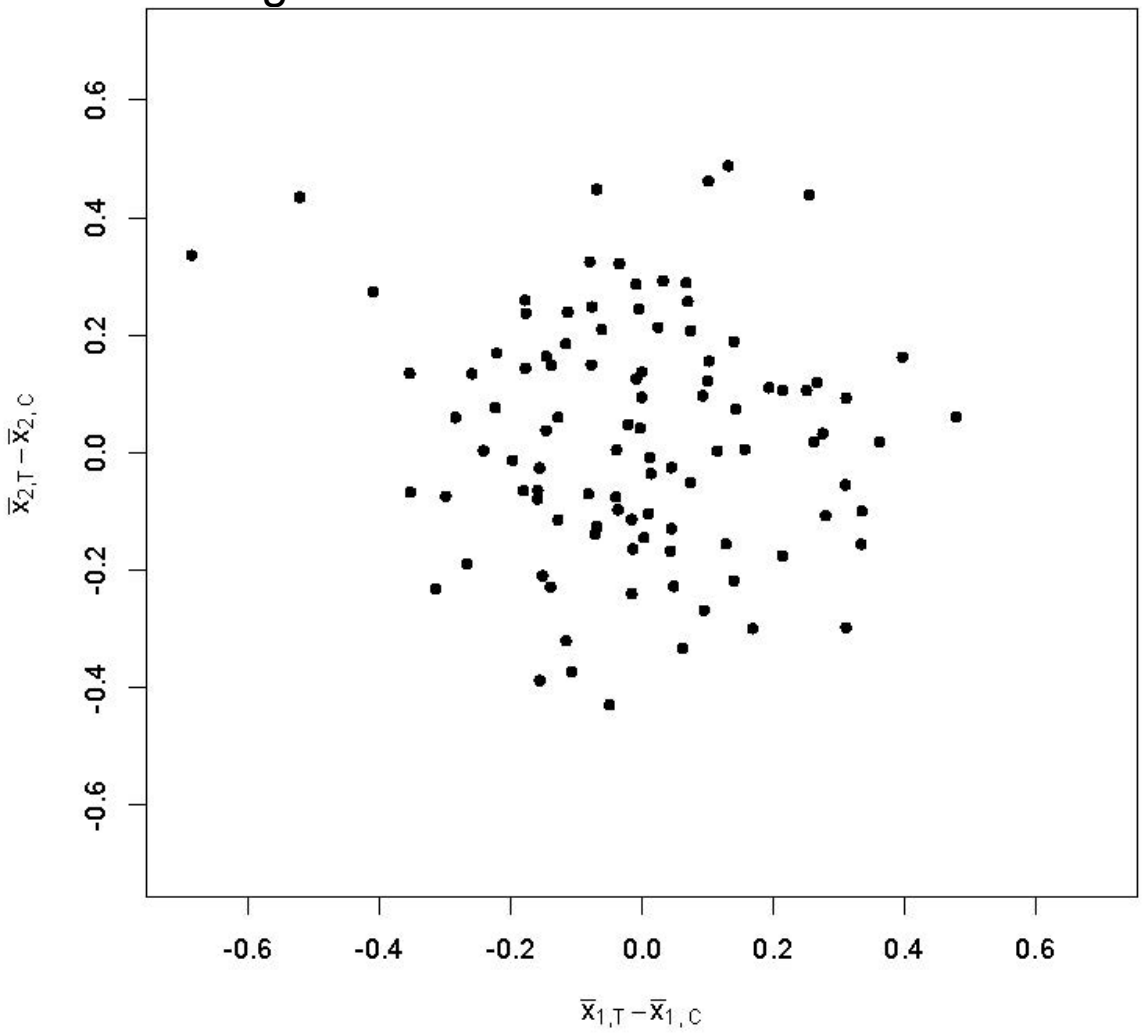


Visualization for two continuous covariates

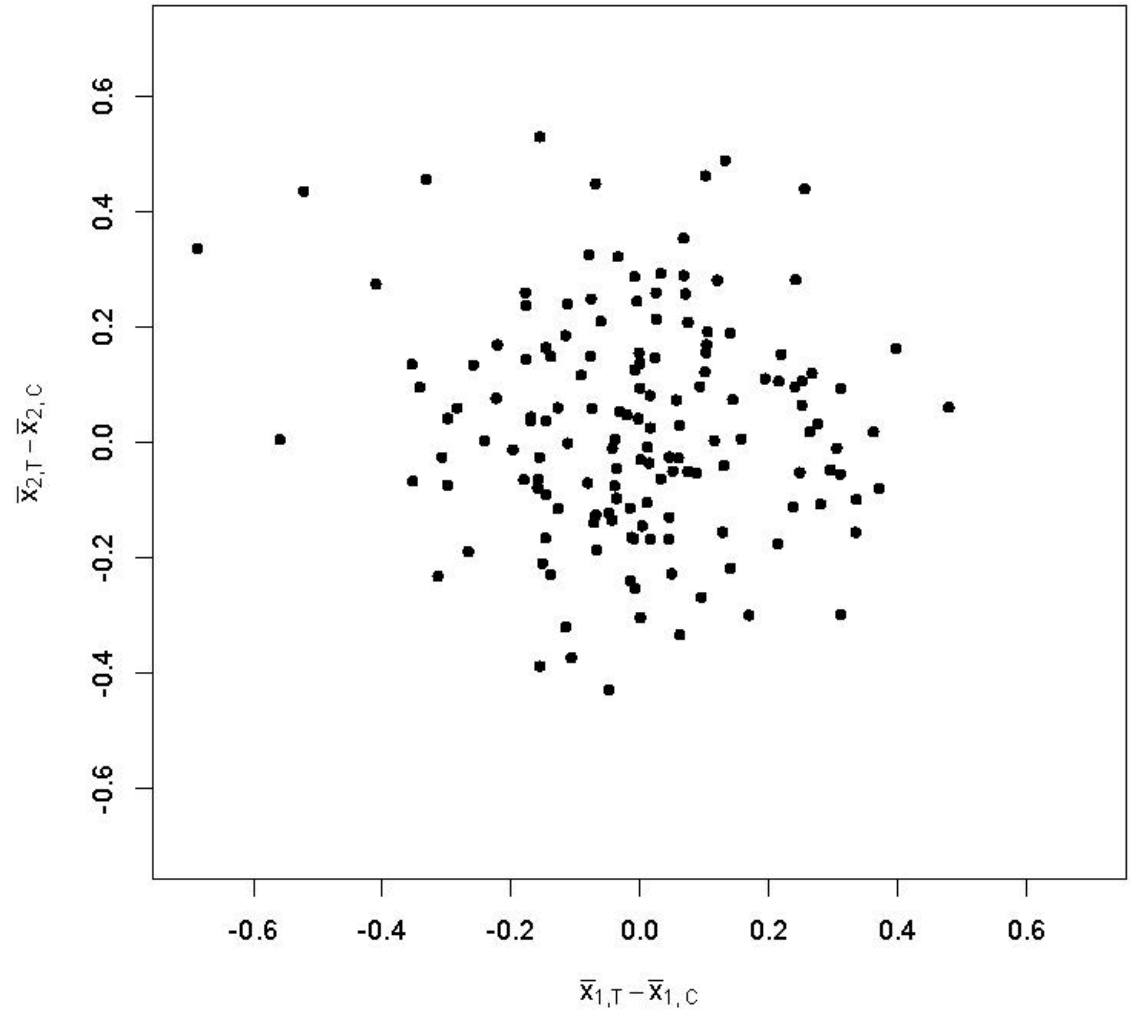


Visualization for two continuous covariates

want relationship between covariates to be close to origin.

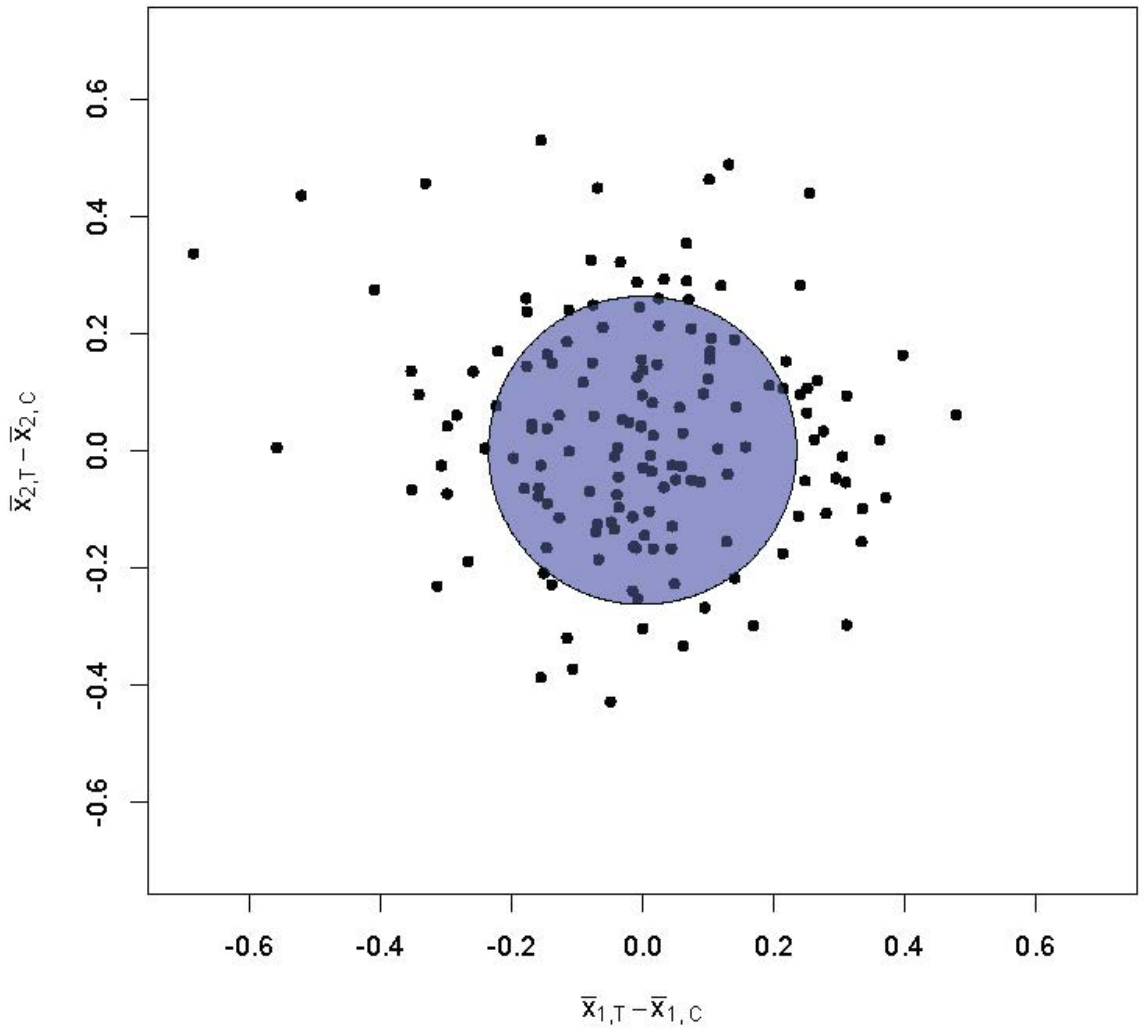


Visualization for two continuous covariates



Visualization for two continuous covariates

accept within a radius from the origin: ie, use a circle



Criterion for re-randomization

- Mahalanobis distance *M* (a multivariate distance between group mean vectors)
  adjusting by variance and covariance of the groups.
- Acceptance criterion: $M \leq a$
- Here *a* is a pre-determined constant
- Trade-off between throwing away randomizations and balancing groups
  circle is small: reject too many?
  circle is large: more likely to be unbalanced (accept further from origin).

Reducing variance of average covariate difference between groups

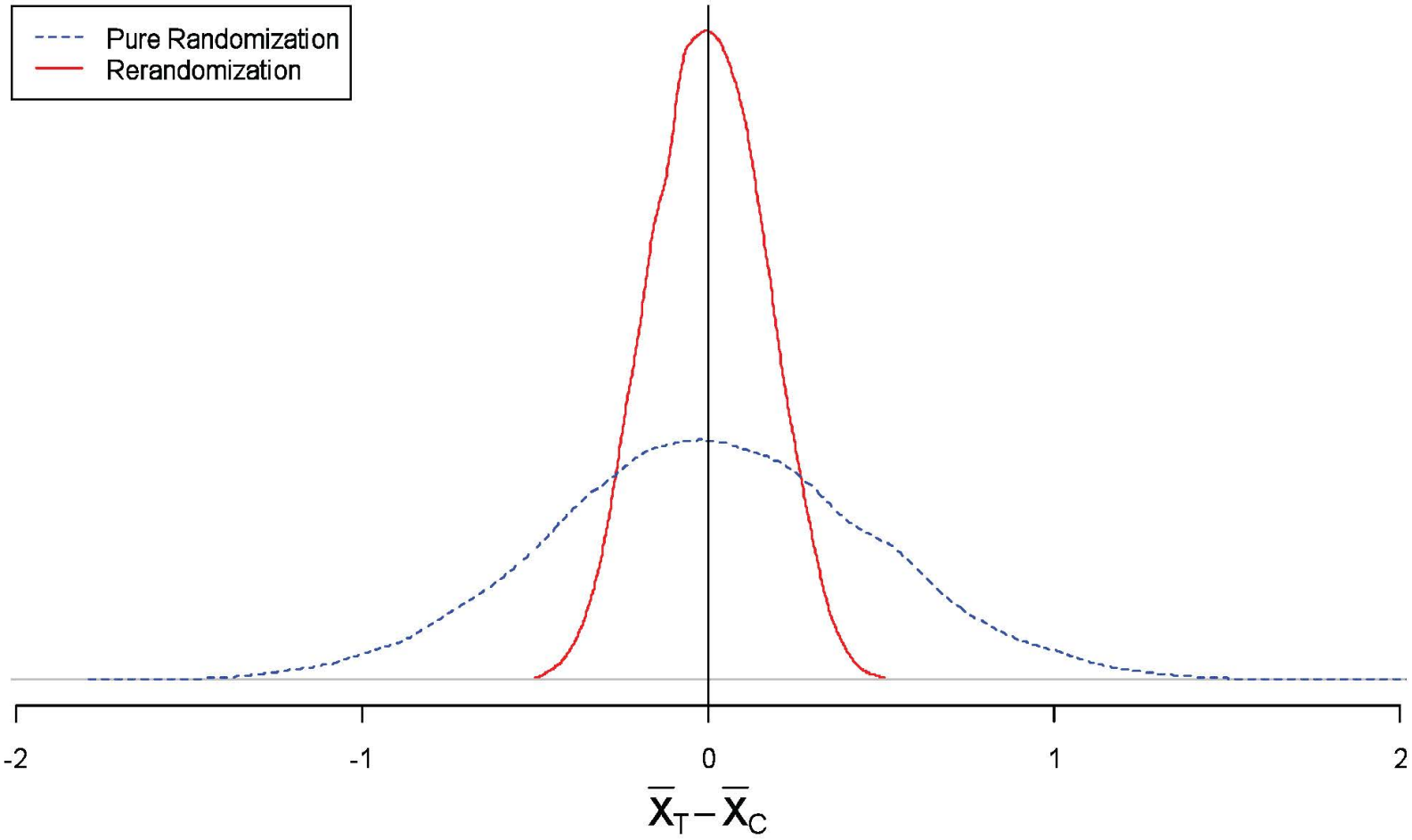


Figure courtesy: Kari Lock Morgan and Donald B. Rubin

Covariate balance achieved by re-randomization - I

bands: how much you shrink distribution of covariates with rerandomization.

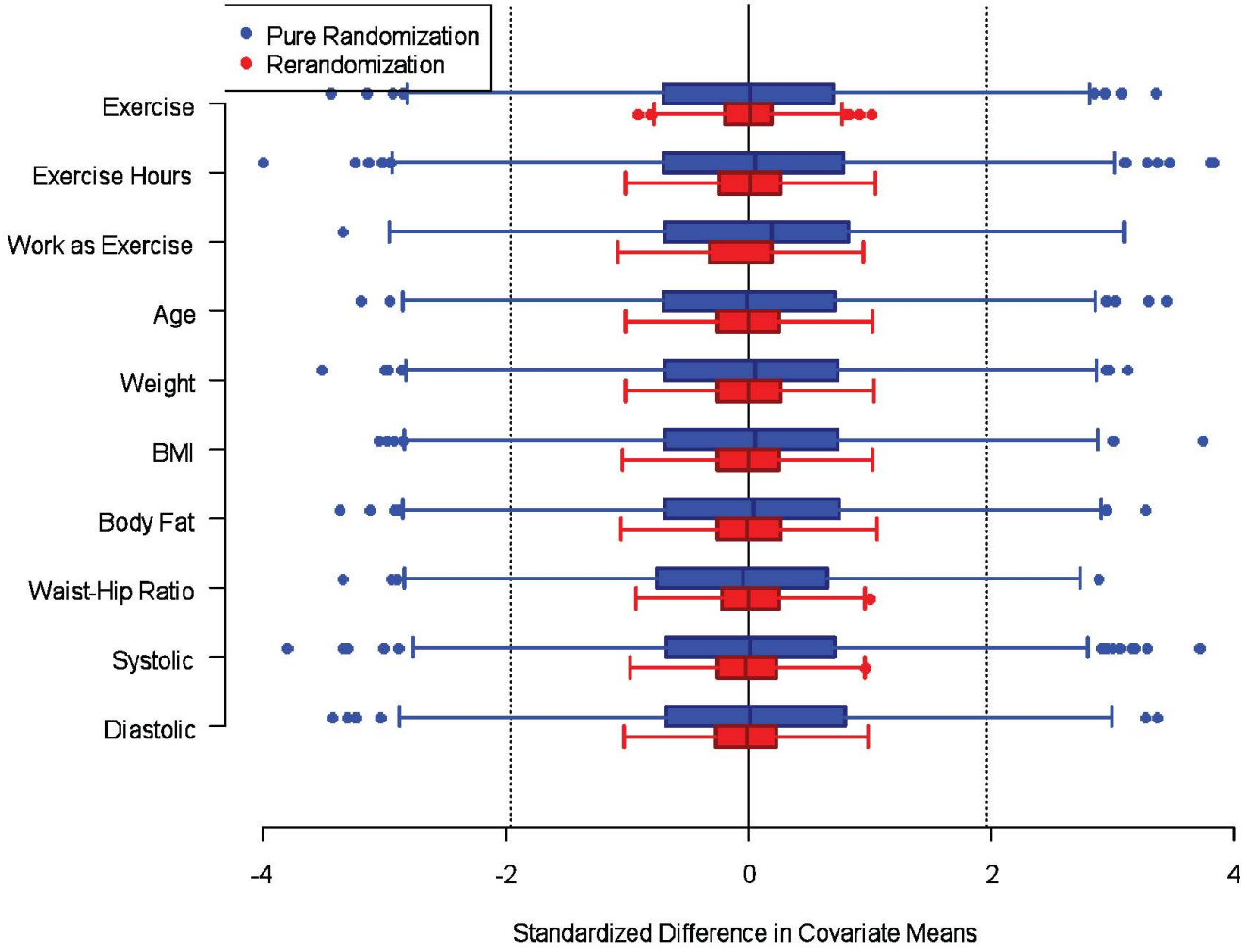
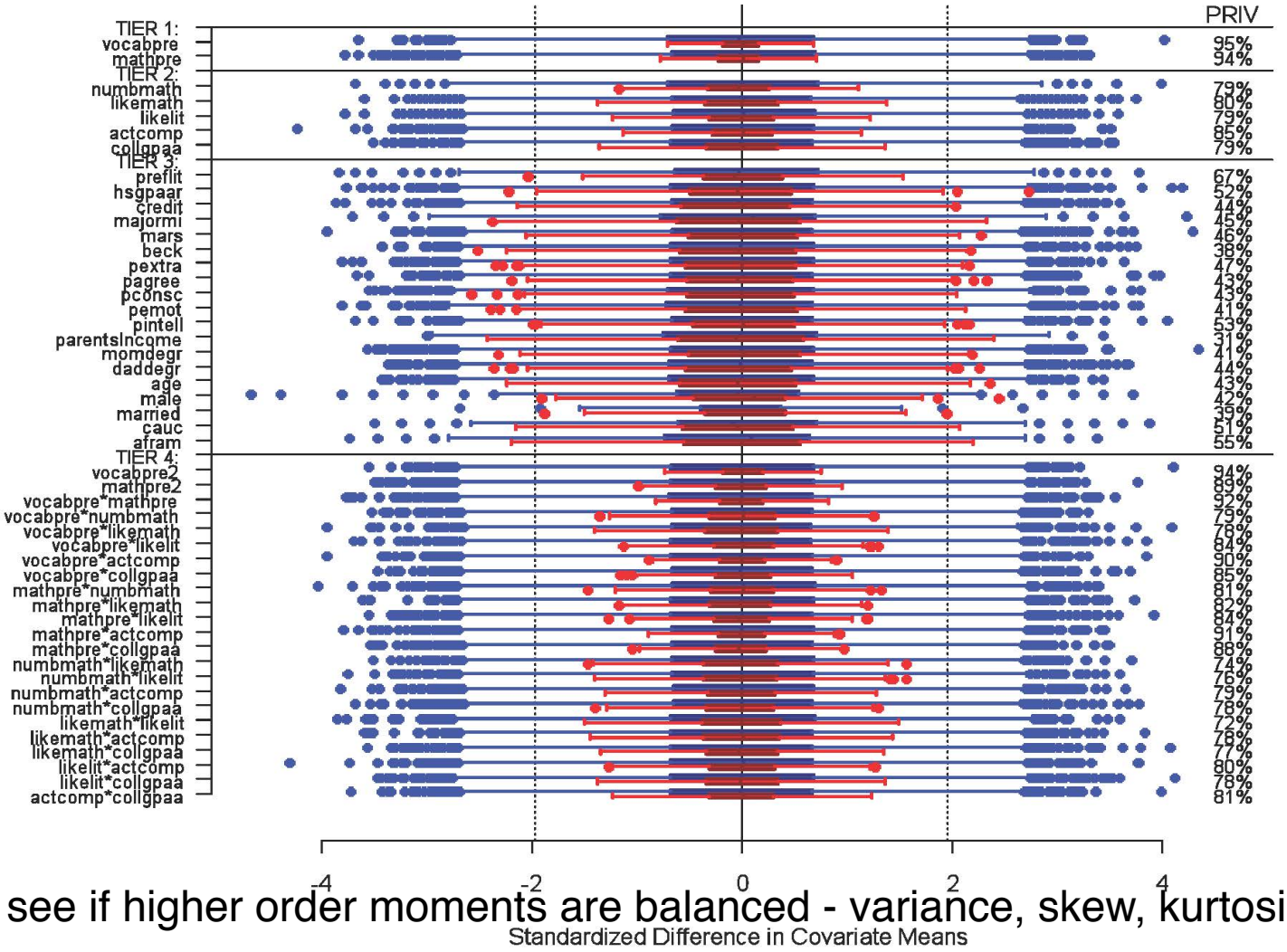


Figure courtesy: Kari Lock Morgan and Donald B. Rubin

need to make choices - requires domain knowledge

Covariate balance achieved by re-randomization - II



Can also see if higher order moments are balanced - variance, skew, kurtosis, etc.

Figure courtesy: Kari Lock Morgan and Donald B. Rubin

Multi-factor experiments

Canabalizing statistical power with this design … too many treatment combinations, I'd say drop some treatments would help in balancing (fewer groups).

- 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), -1 (control).

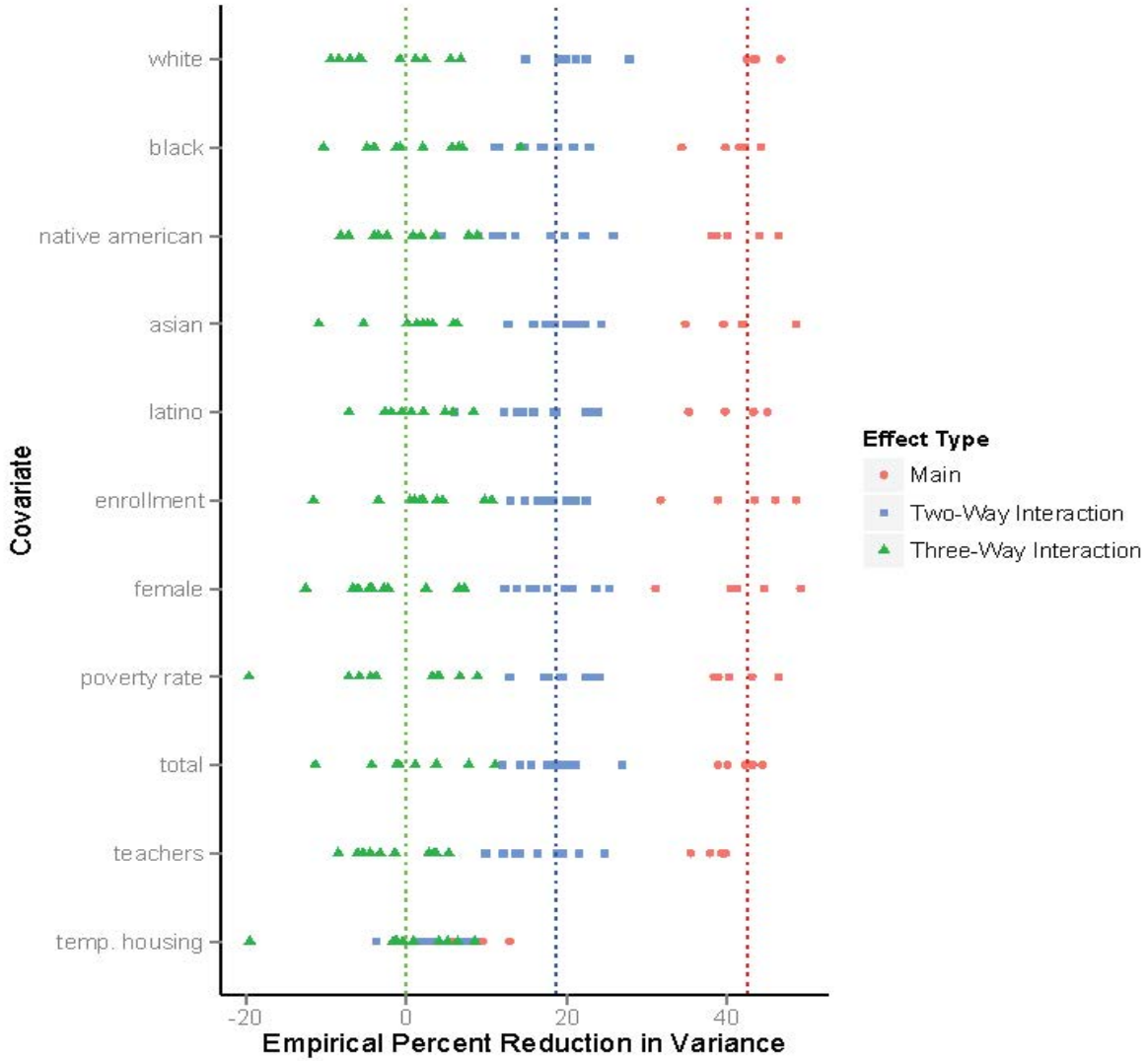
Main effect: effect on margin (on ONE factor: treatment vs control)
Two factor interaction: effect of change in one variable on change in other
Rerandomization strategy: protect main effect most, then 2-factor, 3-factor, etc.

Assignment mechanism

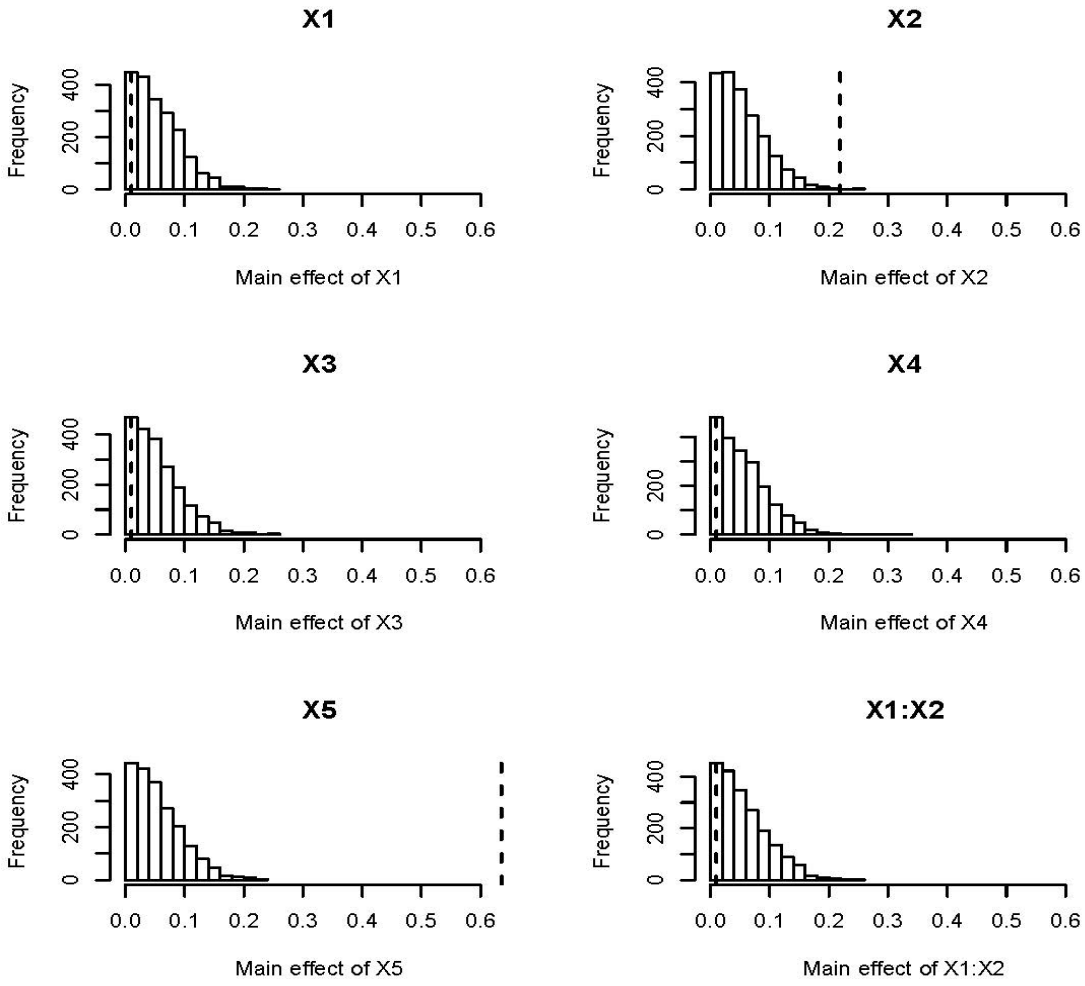
- Completely randomized assignment (CRA) of the 32 treatment combinations to the 224 schools (each treatment to eight schools).
- But need balance over 50 covariates
- Different levels of protection (balance):
  - Maximum protection to five main effects
  - Less protection to two-factor interactions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

Judgement call: domain knowledge.

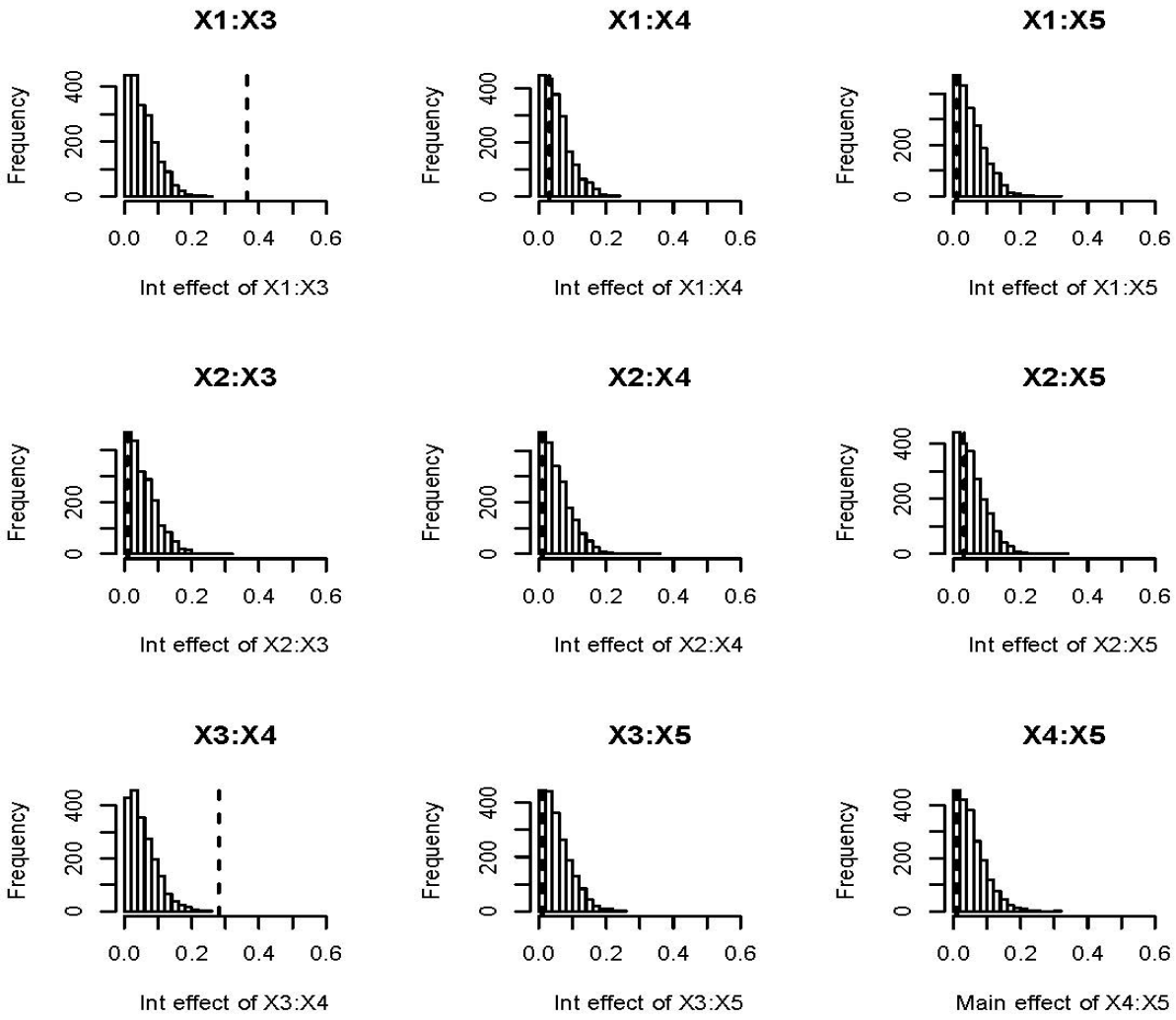
Improving balance by re-randomization



Randomization tests



Randomization tests (contd.)



Modifying Fisher in 2015

- **BLOCK WHAT YOU CAN, RERANDOMIZE WHAT YOU CANNOT!**

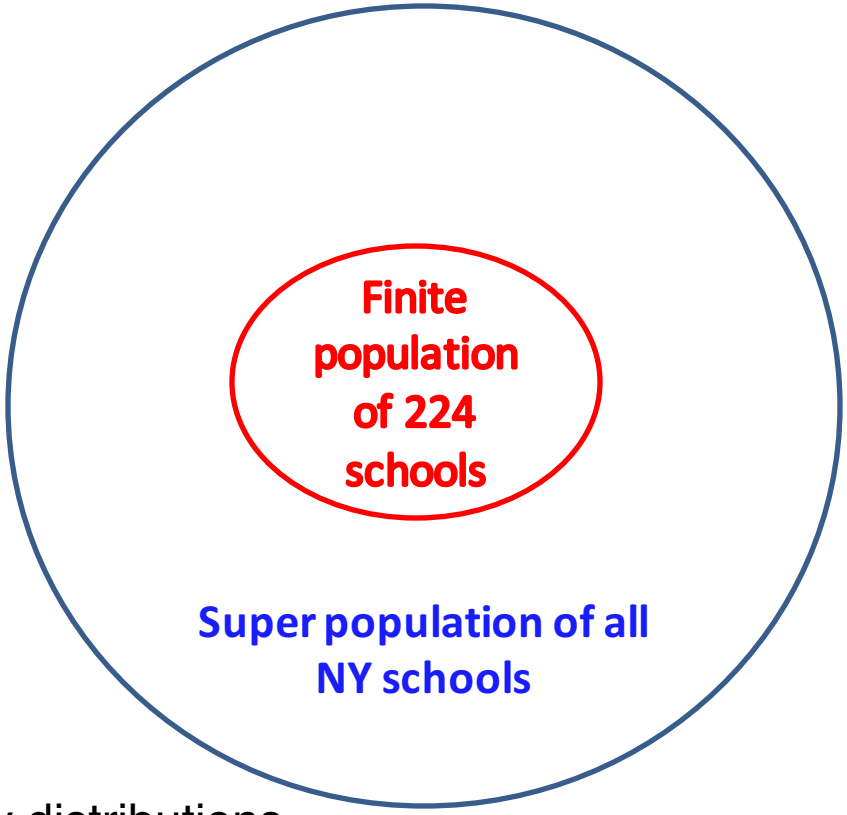
rerandomize specifically: based on a score.

- **ANALYZE AS YOU RERANDOMIZE**

scope of inference: generalizability?
In this lecture: we imagined Y_1, Y_c, Y_t as fixed, our scope was WITHIN sample (ie, not generalizable out of sample).

Fixed to random potential outcomes, Finite to super-population inference

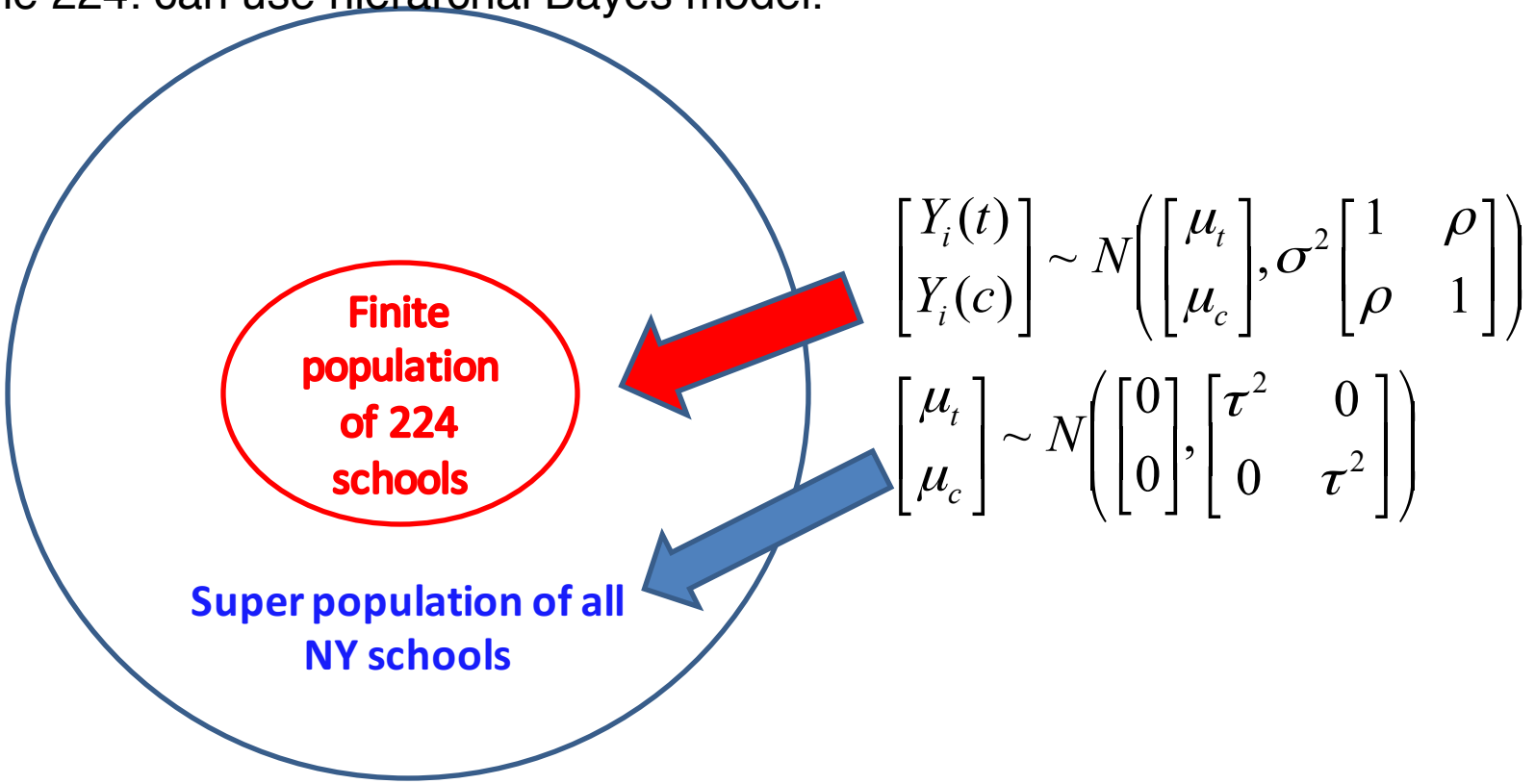
reality: potential outcomes NOT fixed quantities: they are random variables.
can use this to extend to super population —> generalize.



Adding more probability distributions.

Fixed to random potential outcomes, Finite to super-population inference

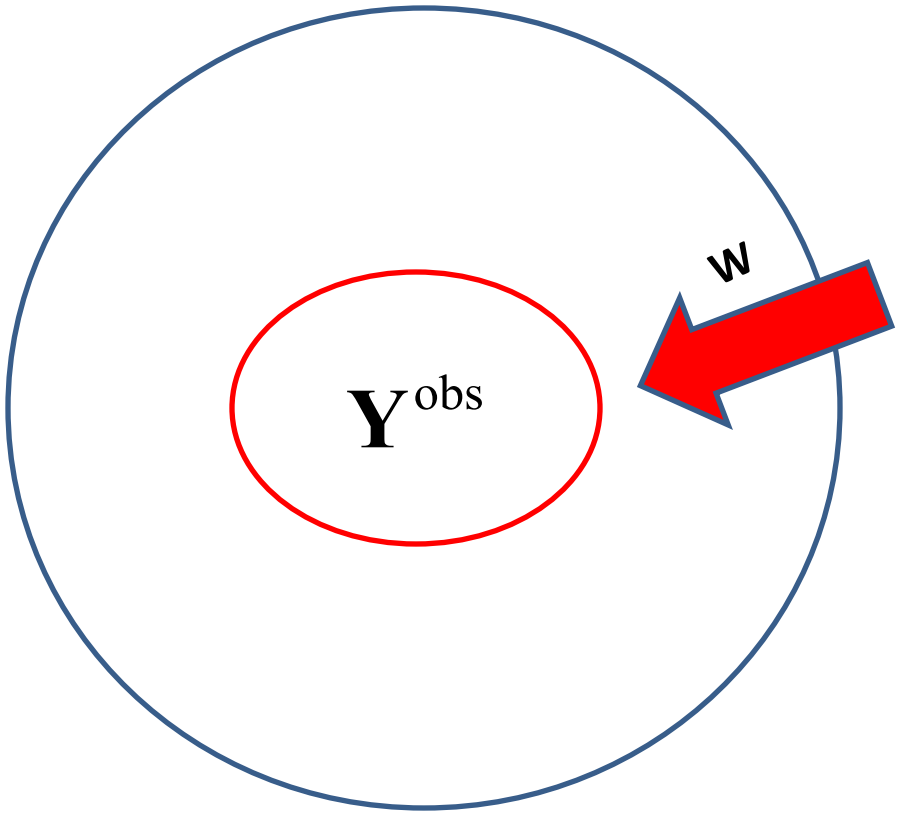
224 units randomly sampled from the super population, and then random sampling within the 224: can use hierarchal Bayes model.



Hierarchical (Bayesian) model

Fixed to random potential outcomes, Finite to super-population inference

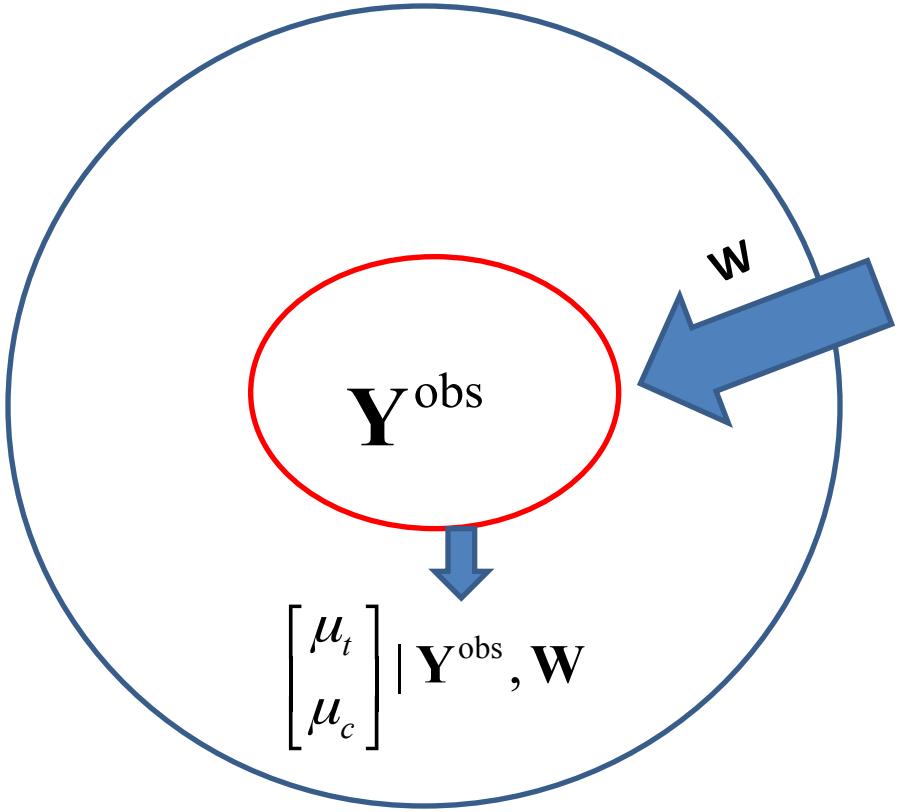
obs for sample of 224 schools.



$$\begin{bmatrix} Y_i(t) \\ Y_i(c) \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_t \\ \mu_c \end{bmatrix}, \sigma^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right)$$

$$\begin{bmatrix} \mu_t \\ \mu_c \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau^2 & 0 \\ 0 & \tau^2 \end{bmatrix} \right)$$

Fixed to random potential outcomes, Finite to super-population inference

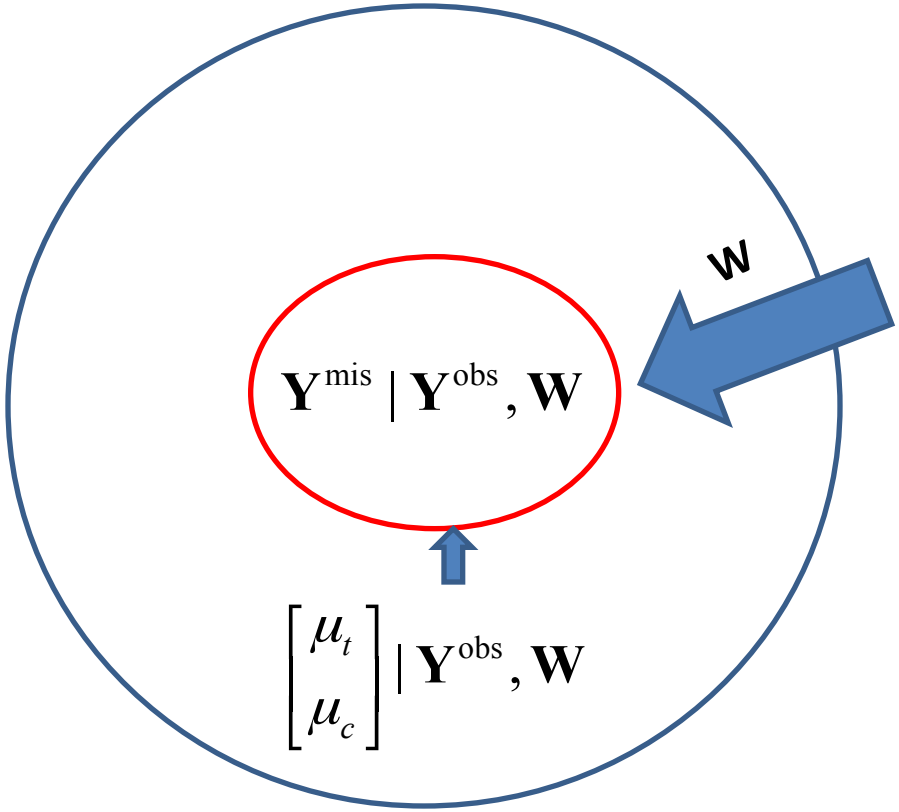


$$\begin{bmatrix} Y_i(t) \\ Y_i(c) \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_t \\ \mu_c \end{bmatrix}, \sigma^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right)$$

$$\begin{bmatrix} \mu_t \\ \mu_c \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau^2 & 0 \\ 0 & \tau^2 \end{bmatrix} \right)$$

Bayes, get posterior distribution on mu's based on observed data and

Back to finite-population inference

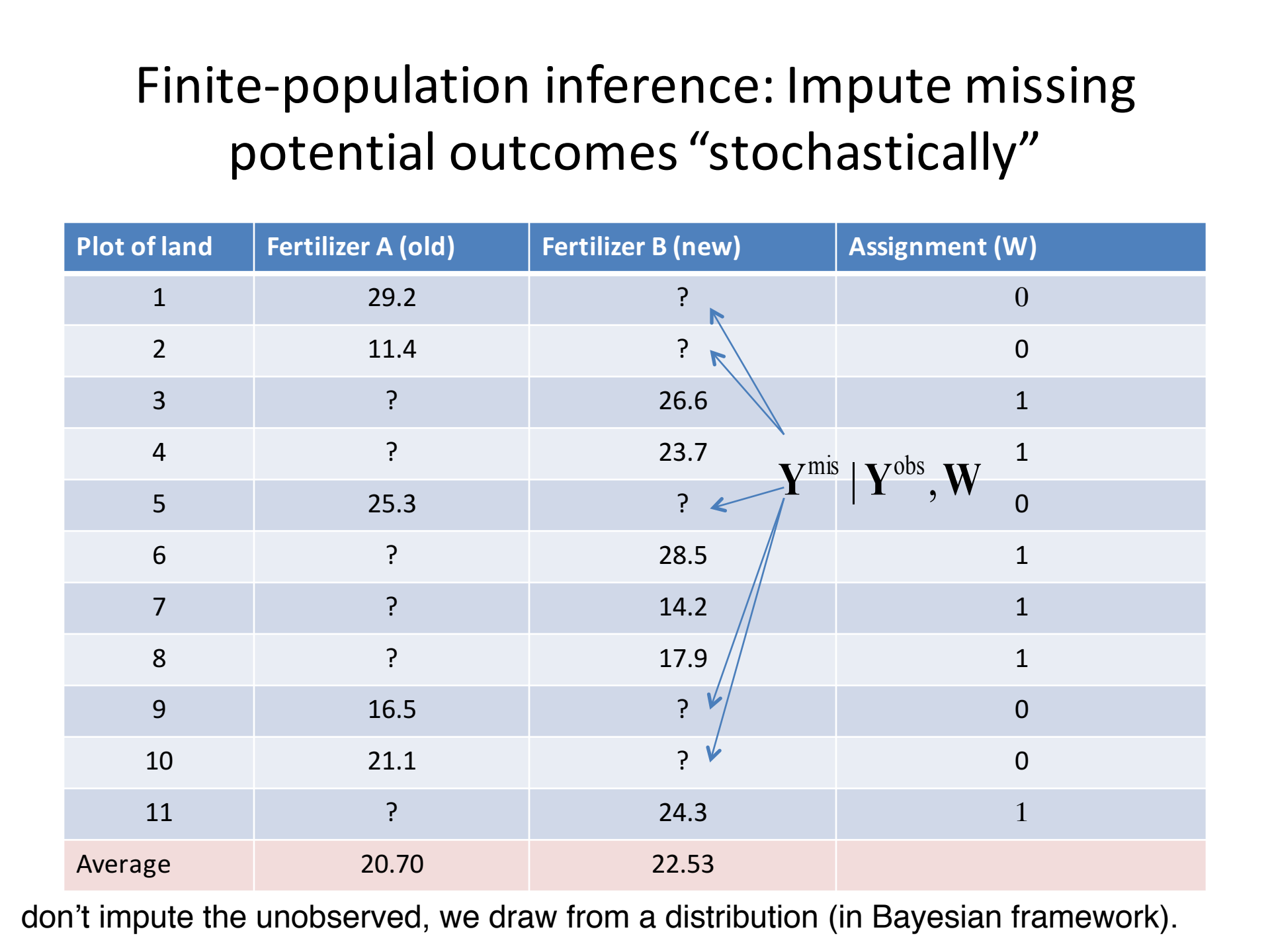


$$\begin{bmatrix} Y_i(t) \\ Y_i(c) \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_t \\ \mu_c \end{bmatrix}, \sigma^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right)$$

$$\begin{bmatrix} \mu_t \\ \mu_c \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau^2 & 0 \\ 0 & \tau^2 \end{bmatrix} \right)$$

Finite-population inference: Impute missing potential outcomes "stochastically"

Plot of land	Fertilizer A (old)	Fertilizer B (new)	Assignment (W)
1	29.2	?	0
2	11.4	?	0
3	?	26.6	1
4	?	23.7	1
5	25.3	?	0
6	?	28.5	1
7	?	14.2	1
8	?	17.9	1
9	16.5	?	0
10	21.1	?	0
11	?	24.3	1
Average	20.70	22.53	



don't impute the unobserved, we draw from a distribution (in Bayesian framework).

To learn more, take Stat 140/240