Why and What to Randomize

BERK ÖZLER

APRIL 25, 2022, WASHINGTON, DC

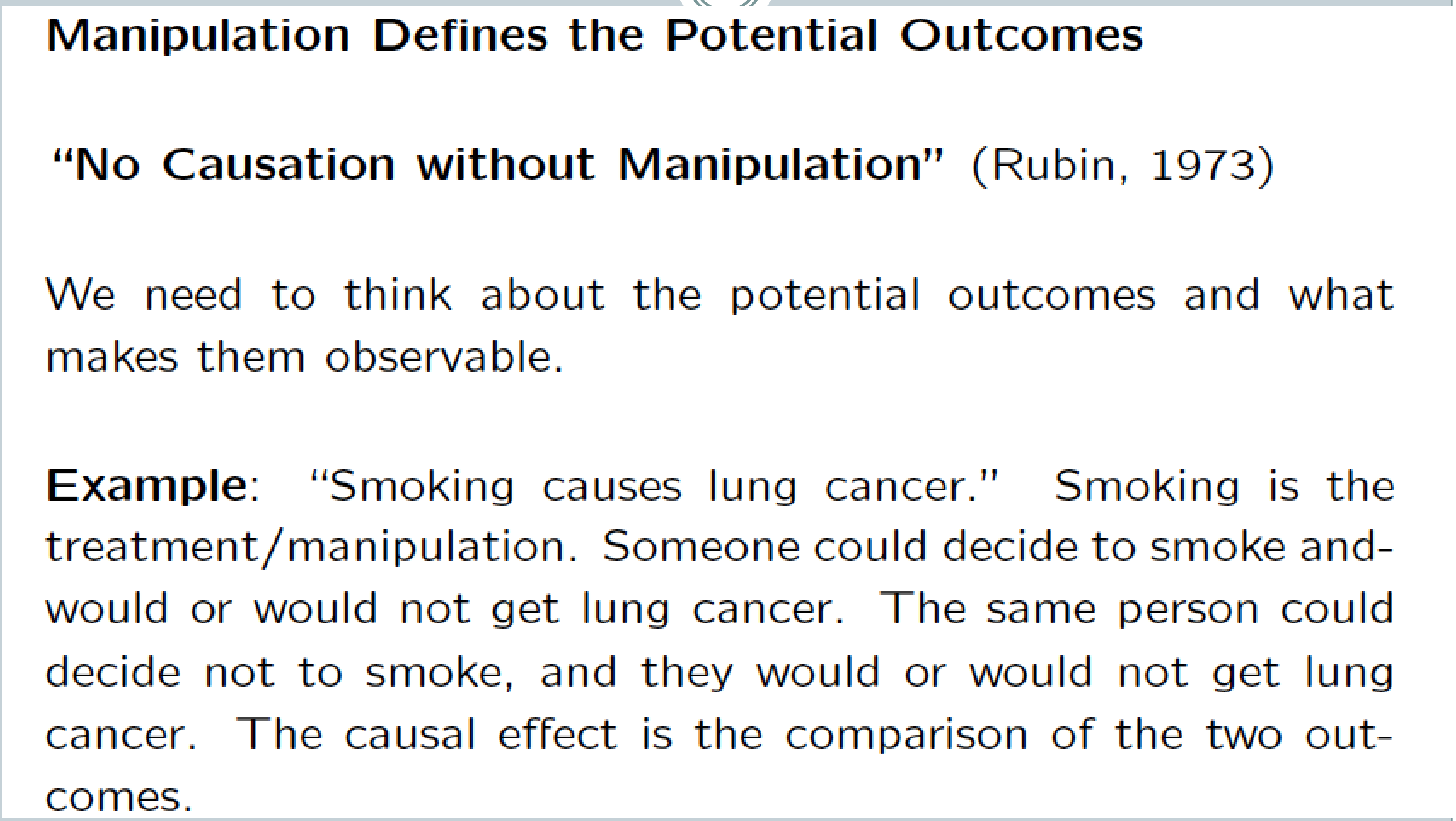
Causality: the Basic Framework

- Potential outcomes and treatments/manipulations
- Unit-level causal effects and causal inference as a "missing data problem"
- Using multiple units and the assumptions needed for estimation of causal effects
- All boiling down to the assignment mechanism…

Causal language

- "My headache went away because I took aspirin."
- "She did not get the job because she is black."
- What do these statements mean? Is the use of "because" causal?
  - The manipulation is clearer for aspirin and no aspirin? What is the manipulation for being black?
- Causality is about "actions" (applied to a unit) and the associated "potential outcomes."

Potential Outcomes



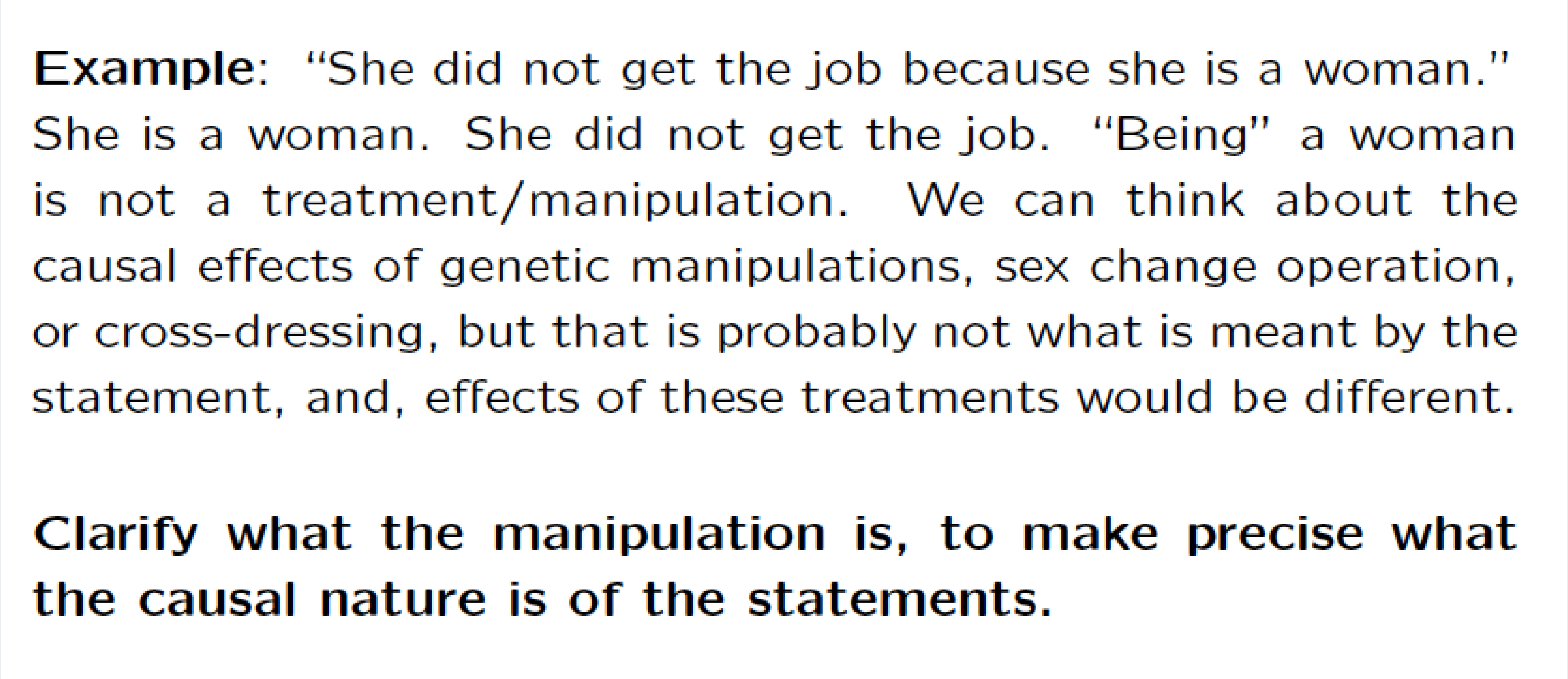
Manipulation Defines the Potential Outcomes

“No Causation without Manipulation” (Rubin, 1973)

We need to think about the potential outcomes and what makes them observable.

Example: “Smoking causes lung cancer.” Smoking is the treatment/manipulation. Someone could decide to smoke and- would or would not get lung cancer. The same person could decide not to smoke, and they would or would not get lung cancer. The causal effect is the comparison of the two outcomes.

Potential Outcomes



Example: "She did not get the job because she is a woman." She is a woman. She did not get the job. "Being" a woman is not a treatment/manipulation. We can think about the causal effects of genetic manipulations, sex change operation, or cross-dressing, but that is probably not what is meant by the statement, and, effects of these treatments would be different.

Clarify what the manipulation is, to make precise what the causal nature is of the statements.

Potential Outcomes

- People do use our definition in real life, including movies…
- “Things would have been much better had I never been born.” (George Bailey, It’s a Wonderful Life)
  - The causal effect of him being born is the entire stream of events in the actual world compared with the counterfactual world without him (that he gets to see thanks to an angel)…
- The “but-for” concept in legal settings.
  - E.g. while calculating damages…

Potential Outcomes

I have a headache. I can take an aspirin. Afterwards my headache may be gone or not.

Unit: "I", Treatment: $W \in \{\text{Asp.}, \text{No Asp.}\}$

Two potential outcomes: state of headache if I take aspirin, state of headache if I dont take aspirin, $Y(\text{Asp.})$, $Y(\text{No Asp.})$.

Causal effect is comparison of those two states for the same unit (me), e.g., $Y(\text{Asp.}) - Y(\text{No Asp.})$.

Fundamental problem of causal inference (Holland, 1986, p. 947): **"it is impossible to observe (both potential outcomes) on the same unit, and therefore we cannot observe the causal effect."**

Potential Outcomes

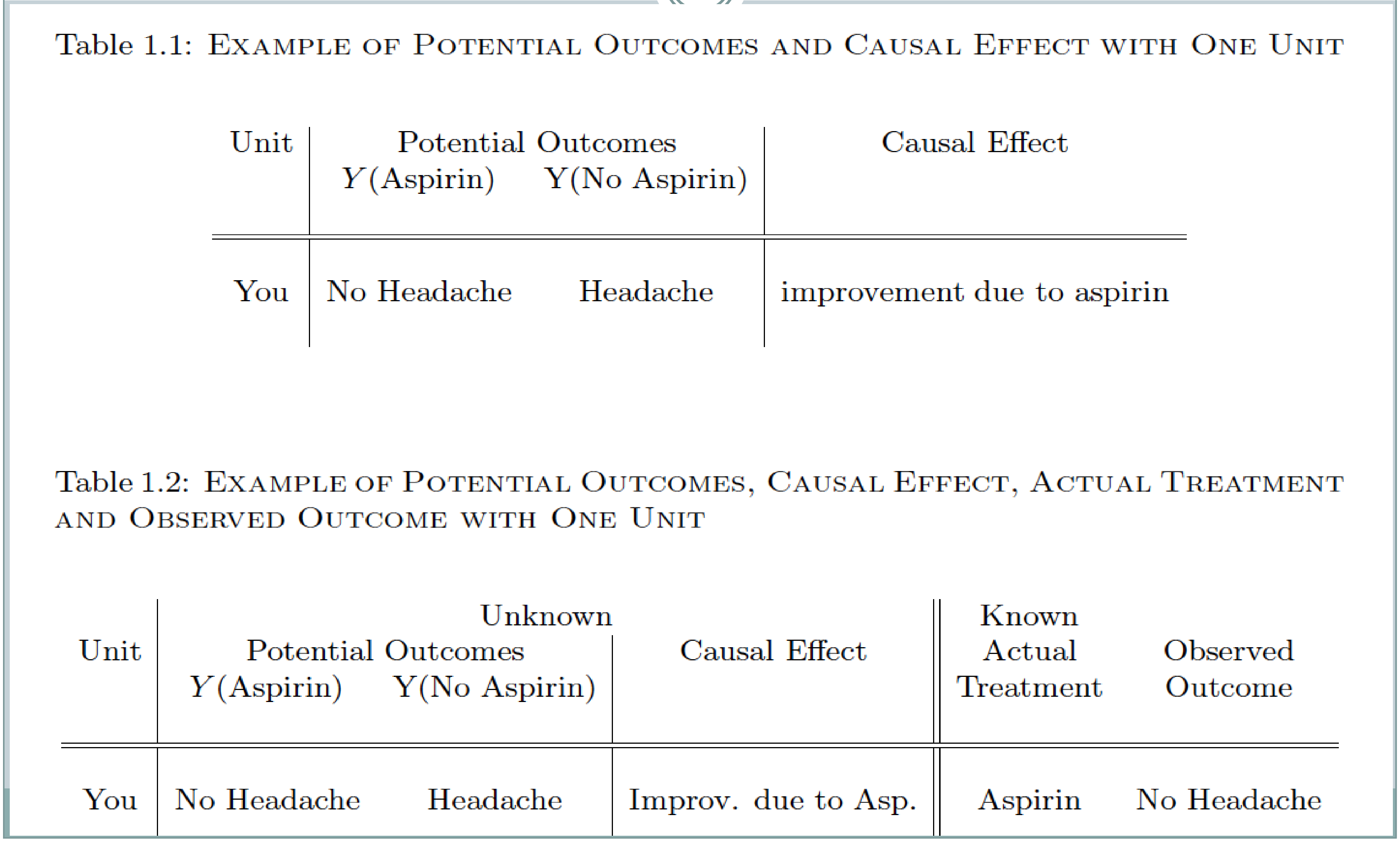


Table 1.1: Example of Potential Outcomes and Causal Effect with One Unit

Unit	Potential Outcomes		Causal Effect
	Y(Aspirin)	Y(No Aspirin)	
You	No Headache	Headache	improvement due to aspirin

Table 1.2: Example of Potential Outcomes, Causal Effect, Actual Treatment and Observed Outcome with One Unit

	Unknown			Known	
Unit	Potential Outcomes		Causal Effect	Actual	Observed
	Y(Aspirin)	Y(No Aspirin)		Treatment	Outcome
You	No Headache	Headache	Improv. due to Asp.	Aspirin	No Headache

Causal inference as a “missing data” problem

- Given any treatment assigned to an individual unit, the outcome(s) associated with any alternative treatment(s) is missing!
    - We observe at most half of the potential outcomes and **none** of the unit-level causal effects.
- Statements about individual outcomes then pose philosophical problems:
    - “If he had taken that new drug, he would not have died so soon.”
        - This may be an expert opinion (and perhaps a real good one), but it is not a causal statement.

Causal inference as a "missing data" problem

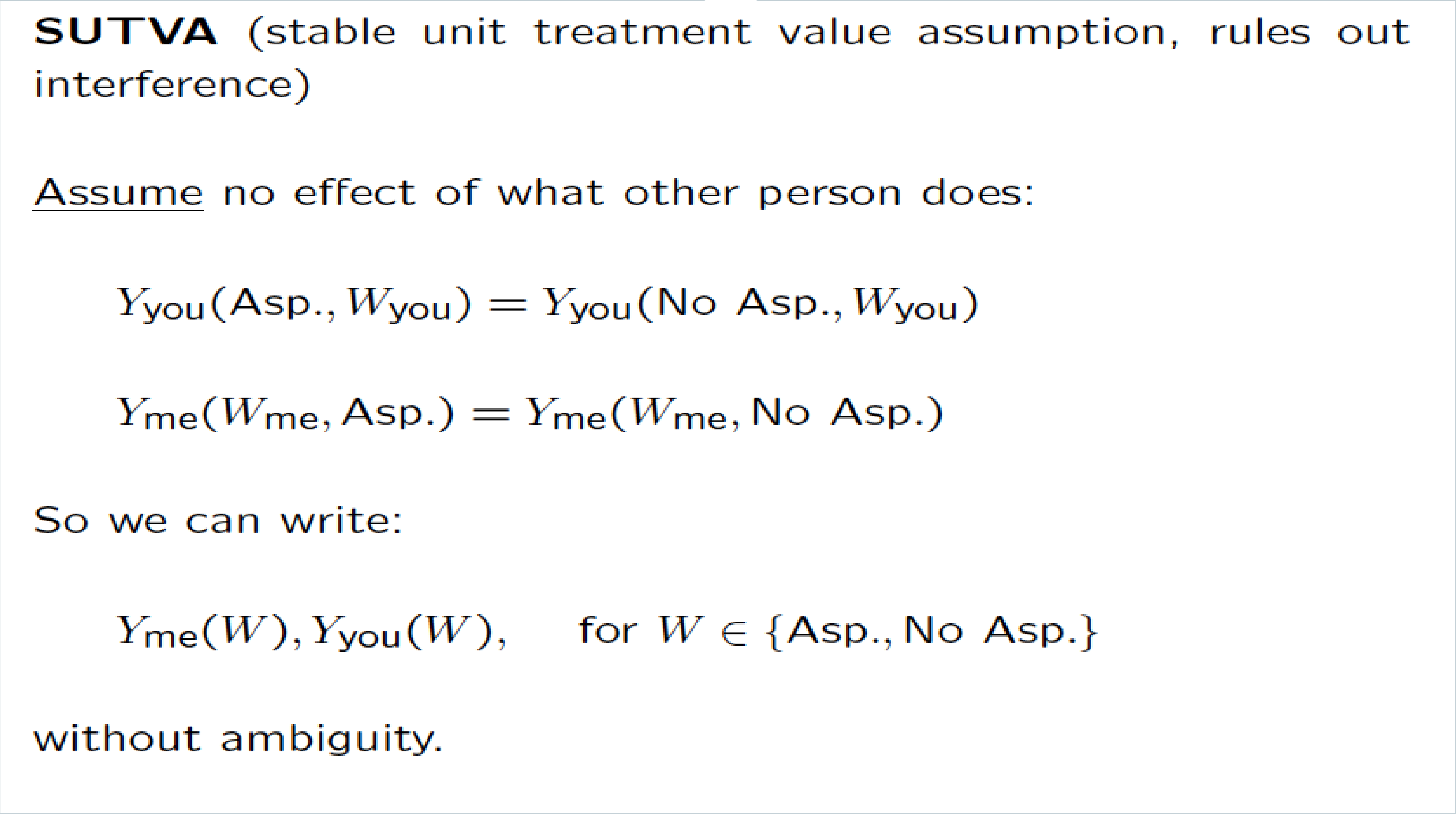
- Given that only one potential outcome can be observed for any one person, there is a need to observe multiple units to be able to conduct causal inference.
  - Notice that we can define causal impacts with respect to actions and potential outcomes for a single unit, but we need multiple units for estimation.
- For this purpose "you today" and "you tomorrow" are two different units.
  - You may develop sensitivity (or lack thereof) to aspirin over time; AM may be different than PM; intensity of headache may vary, etc.
  - More common is "you today" vs. "me today"

Use of multiple units to estimate causal effects

- The use of multiple units, however, does not come close to solving the problem. Let's start with the multiplicity of potential outcomes:

- Suppose now we have two units: "me" and "you." We can each take aspirin or not. Now there are four potential outcomes for each of us: the state of my headache in a 2x2 matrix of "me/you" X "aspirin/no aspirin."
  - Notice that you can ignore this "dependence" but you MUST notice that it is an assumption that might be wrong!

Stable Unit Treatment Value Assumption (SUTVA)



SUTVA (stable unit treatment value assumption, rules out interference)

Assume no effect of what other person does:

$$Y_{\text{you}}(\text{Asp.}, W_{\text{you}}) = Y_{\text{you}}(\text{No Asp.}, W_{\text{you}})$$

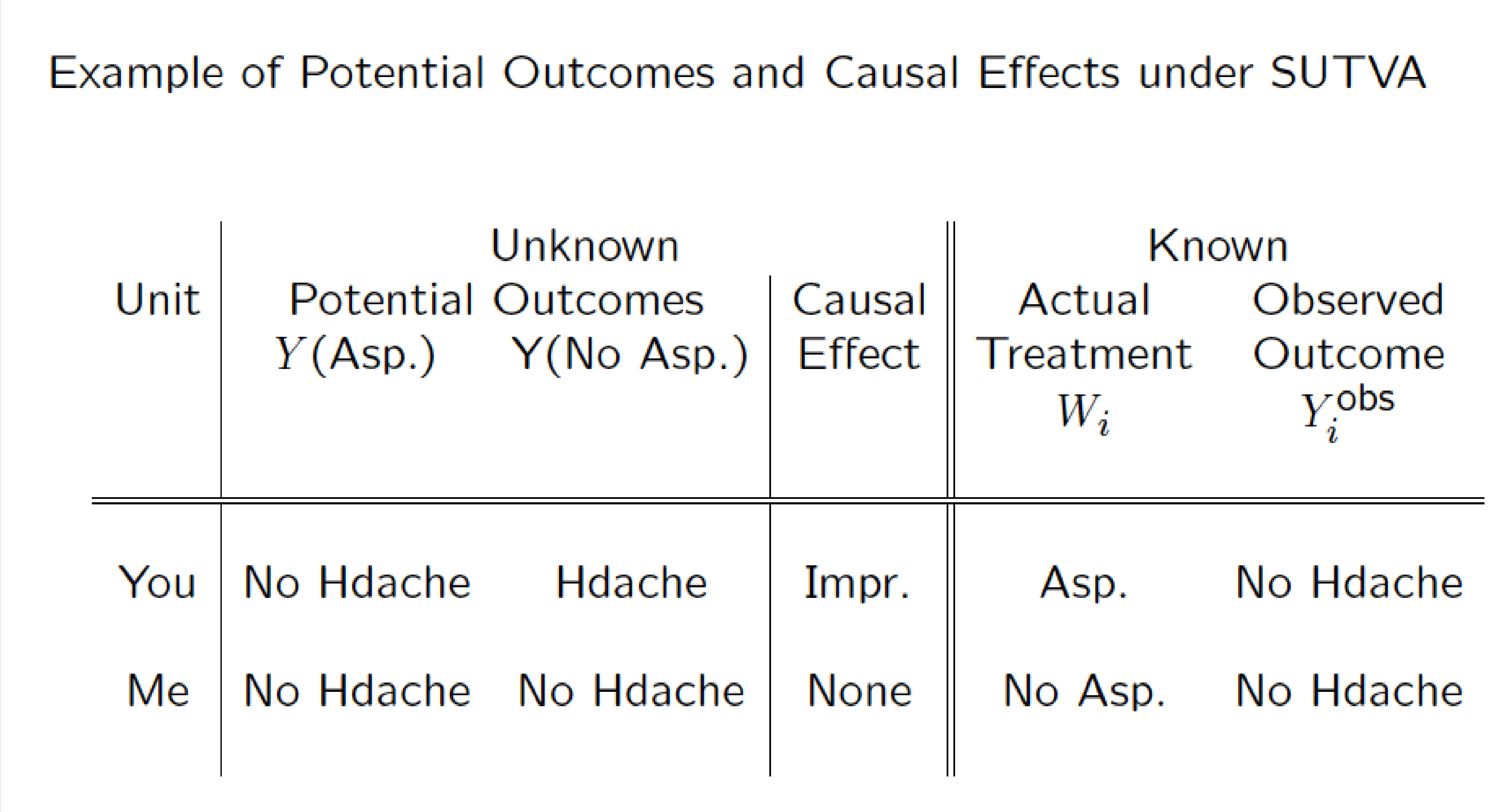
$$Y_{\text{me}}(W_{\text{me}}, \text{Asp.}) = Y_{\text{me}}(W_{\text{me}}, \text{No Asp.})$$

So we can write:

$$Y_{\text{me}}(W), Y_{\text{you}}(W), \quad \text{for } W \in \{\text{Asp.}, \text{No Asp.}\}$$

without ambiguity.

Stable Unit Treatment Value Assumption (SUTVA)



Example of Potential Outcomes and Causal Effects under SUTVA

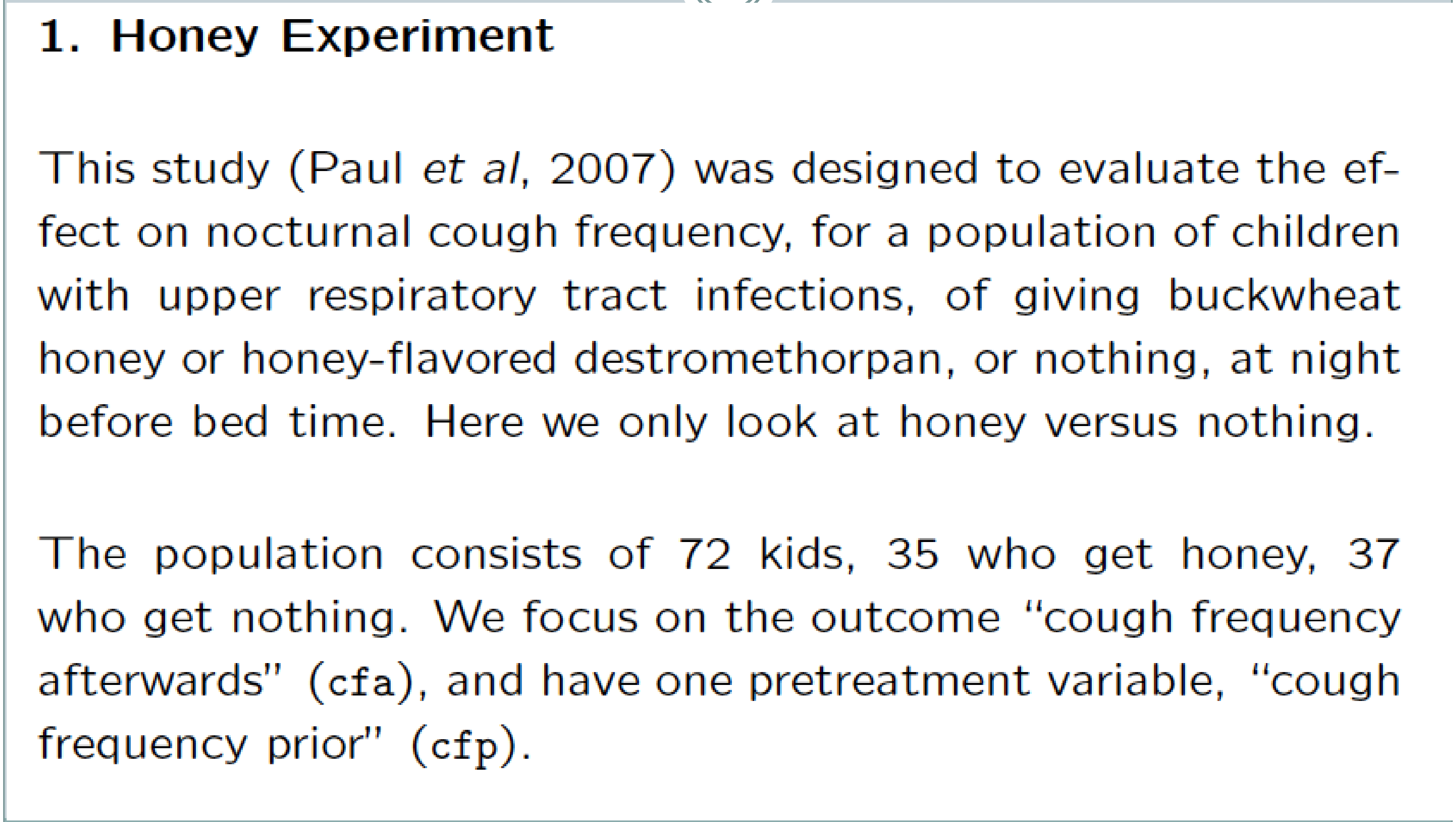
	Unknown			Known	
Unit	Potential Outcomes		Causal	Actual	Observed
	Y(Asp.)	Y(No Asp.)	Effect	Treatment	Outcome
				W_i	Y_i^{obs}
You	No Hdache	Hdache	Impr.	Asp.	No Hdache
Me	No Hdache	No Hdache	None	No Asp.	No Hdache

Stable Unit Treatment Value Assumption (SUTVA)

Most of the time we will assume that the potential outcomes are indexed only by the treatment received by that unit, not by treatments received by other units. Strong assumption.

- guard rows in agricultural experiments
- Infectious diseases: vaccination for one person affects outcomes for other individuals.
- General equilibrium effects: providing job training for some may affect labor market prospects for others.
- Teaching some students in a class may affect others.

Filling in the "missing values"

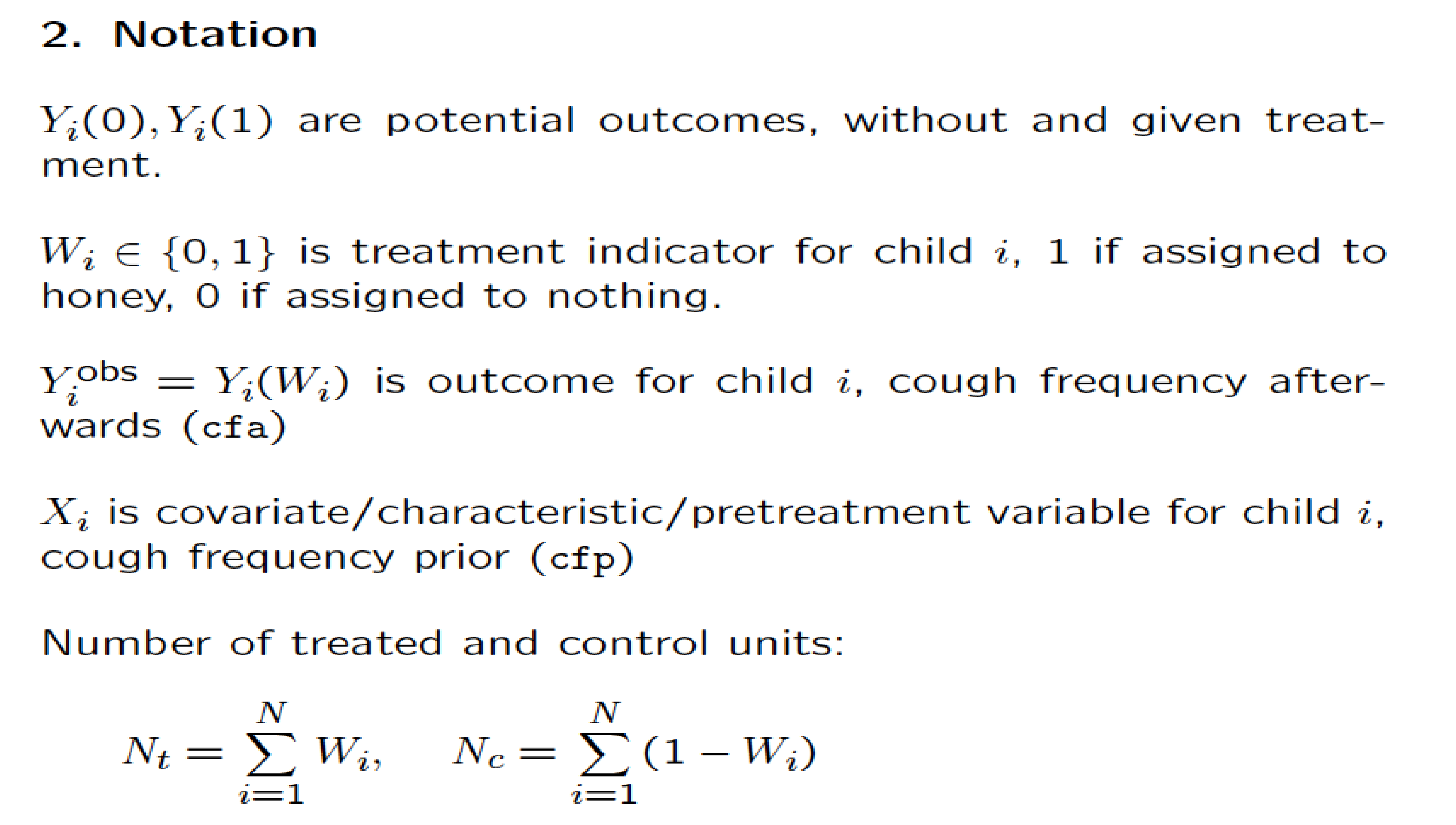


1. Honey Experiment

This study (Paul *et al*, 2007) was designed to evaluate the effect on nocturnal cough frequency, for a population of children with upper respiratory tract infections, of giving buckwheat honey or honey-flavored destromethorpan, or nothing, at night before bed time. Here we only look at honey versus nothing.

The population consists of 72 kids, 35 who get honey, 37 who get nothing. We focus on the outcome "cough frequency afterwards" (`cfa`), and have one pretreatment variable, "cough frequency prior" (`cfp`).

Filling in the "missing values"



2. Notation

$Y_i(0), Y_i(1)$ are potential outcomes, without and given treatment.

$W_i \in \{0, 1\}$ is treatment indicator for child i, 1 if assigned to honey, 0 if assigned to nothing.

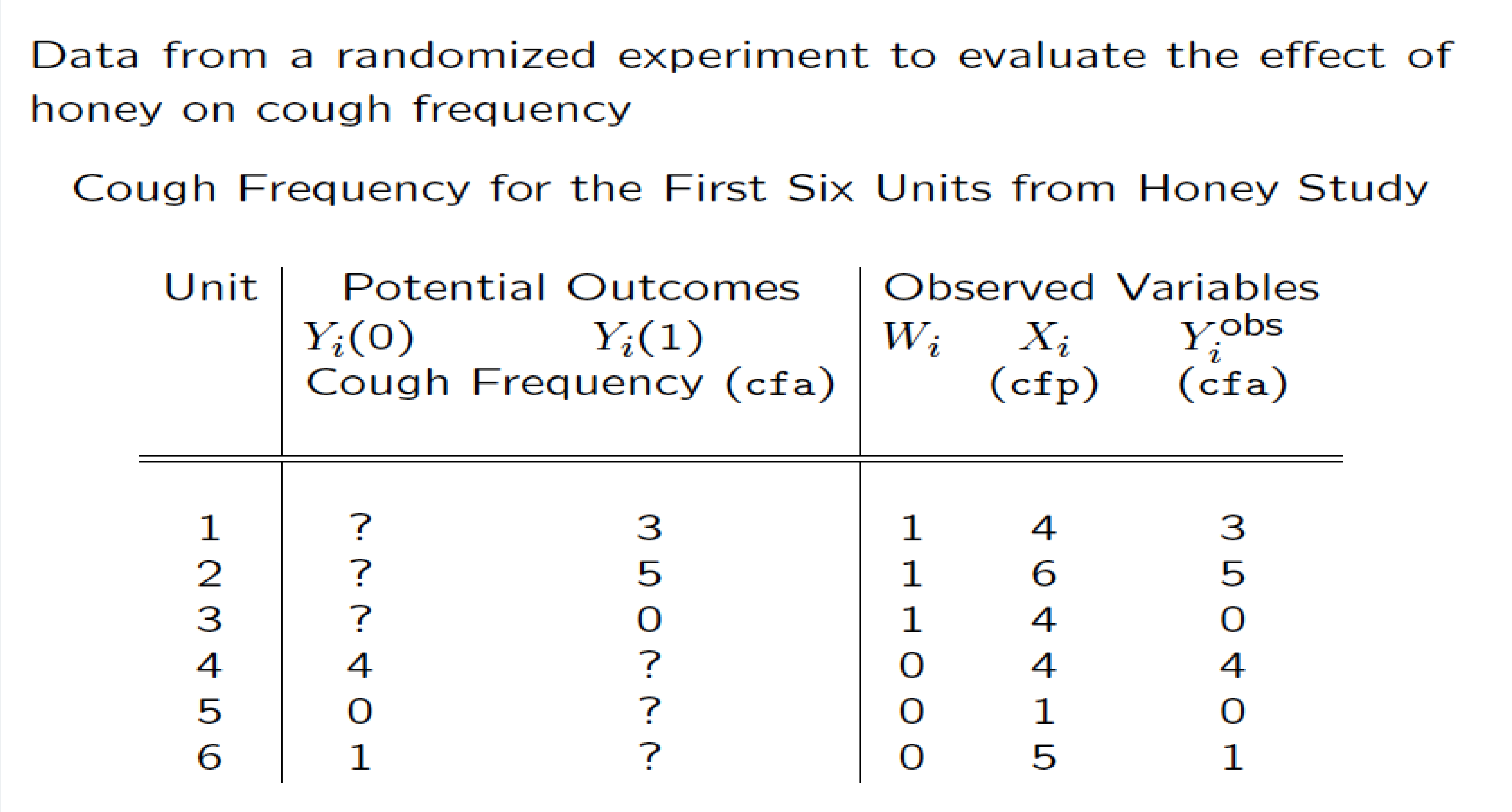
$Y_i^{\text{obs}} = Y_i(W_i)$ is outcome for child i, cough frequency afterwards (`cfa`)

X_i is covariate/characteristic/pretreatment variable for child i, cough frequency prior (`cfp`)

Number of treated and control units:

$$N_t = \sum_{i=1}^{N} W_i, \qquad N_c = \sum_{i=1}^{N} (1 - W_i)$$

Filling in the “missing values”



Data from a randomized experiment to evaluate the effect of honey on cough frequency

Cough Frequency for the First Six Units from Honey Study

Unit	Potential Outcomes		Observed Variables		
	$Y_i(0)$	$Y_i(1)$	W_i	X_i	Y_i^{obs}
	Cough Frequency (cfa)			(cfp)	(cfa)
1	?	3	1	4	3
2	?	5	1	6	5
3	?	0	1	4	0
4	4	?	0	4	4
5	0	?	0	1	0
6	1	?	0	5	1

Filling in the “missing values”

3. Fisher's Approach

Assess a null hypothesis:

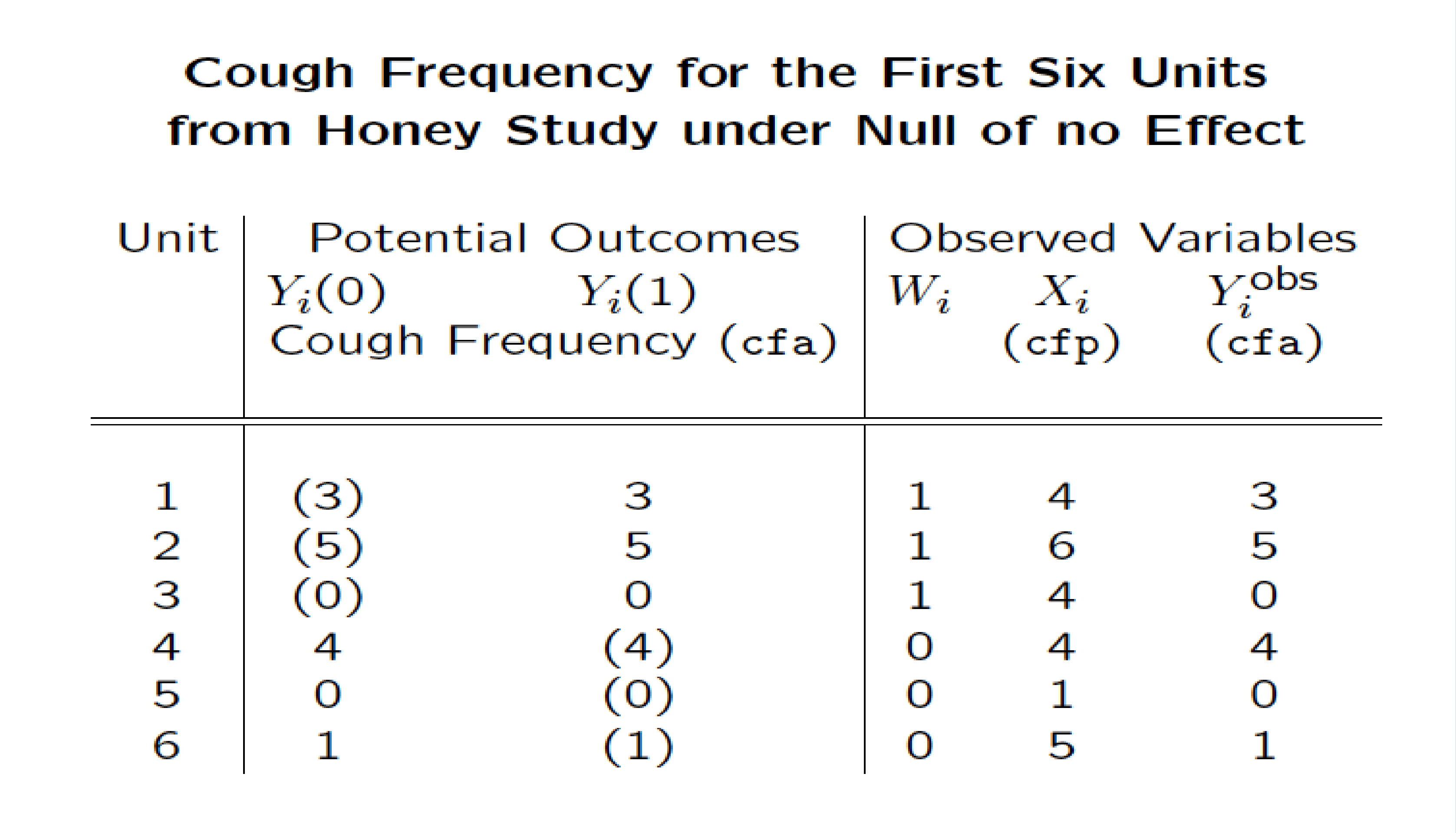
$$H_0 : \ Y_i(0) = Y_i(1) \text{ for all } i = 1, \ldots, N$$

against the alternative that for some units there is some effect of the treatment.

Key Feature

The null hypothesis is **sharp**: under the null hypothesis we know everything, we can fill in all the missing potential outcomes in the table.

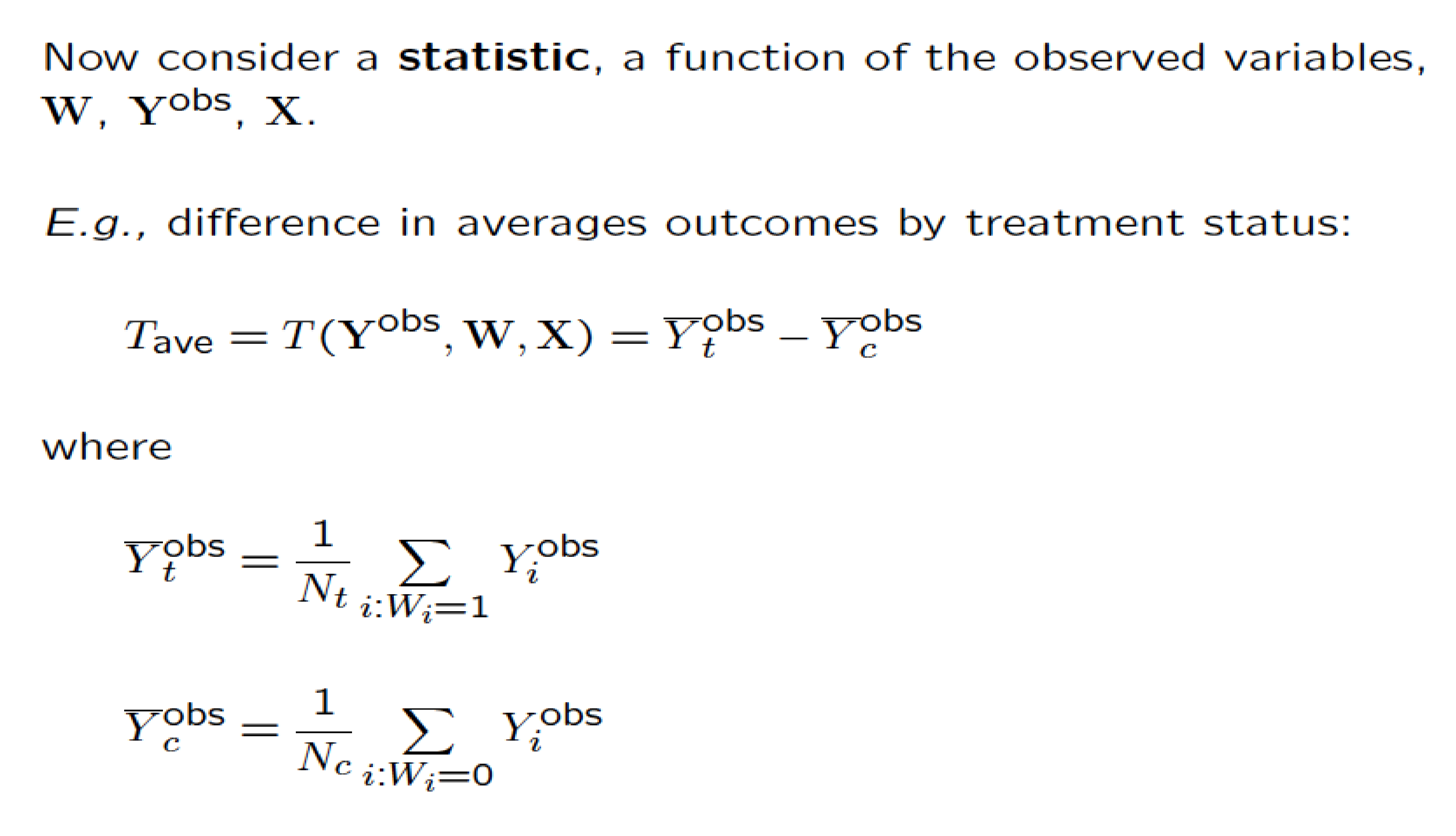
Filling in the “missing values”



Cough Frequency for the First Six Units from Honey Study under Null of no Effect

Unit	Potential Outcomes		Observed Variables		
	$Y_i(0)$	$Y_i(1)$	W_i	X_i	Y_i^{obs}
	Cough Frequency (cfa)			(cfp)	(cfa)
1	(3)	3	1	4	3
2	(5)	5	1	6	5
3	(0)	0	1	4	0
4	4	(4)	0	4	4
5	0	(0)	0	1	0
6	1	(1)	0	5	1

Filling in the "missing values"



Now consider a **statistic**, a function of the observed variables, $\mathbf{W}$, $\mathbf{Y}^{\text{obs}}$, $\mathbf{X}$.

E.g., difference in averages outcomes by treatment status:

$$T_{\text{ave}} = T(\mathbf{Y}^{\text{obs}}, \mathbf{W}, \mathbf{X}) = \overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}}$$

where

$$\overline{Y}_t^{\text{obs}} = \frac{1}{N_t} \sum_{i:W_i=1} Y_i^{\text{obs}}$$

$$\overline{Y}_c^{\text{obs}} = \frac{1}{N_c} \sum_{i:W_i=0} Y_i^{\text{obs}}$$

Filling in the "missing values"

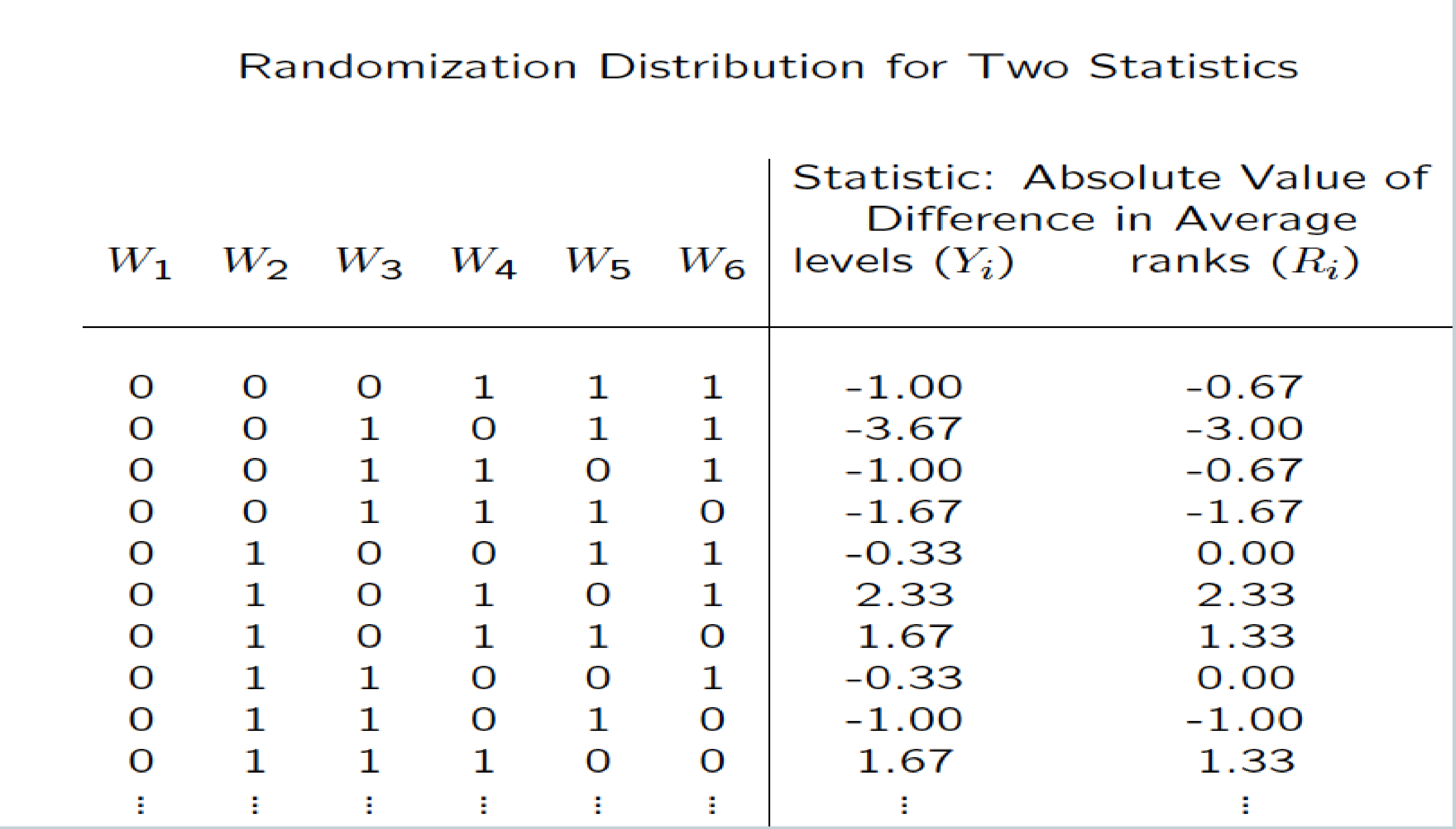
Given our sample of six units, the value of the statistic is

$$T_{\text{ave}} = \frac{8}{3} - \frac{5}{3} = 1$$

Fisher wants to assess how unusual this value of $T_{\text{ave}} = 1$ is, under the null hypothesis that there is no effect of the treatment whatsoever.

The key insight is that we can derive the *exact* distribution of $T(\mathbf{Y}^{\text{obs}}, \mathbf{W}, \mathbf{X})$ under the randomization distribution (the distribution induced by random assignment to the treatment).

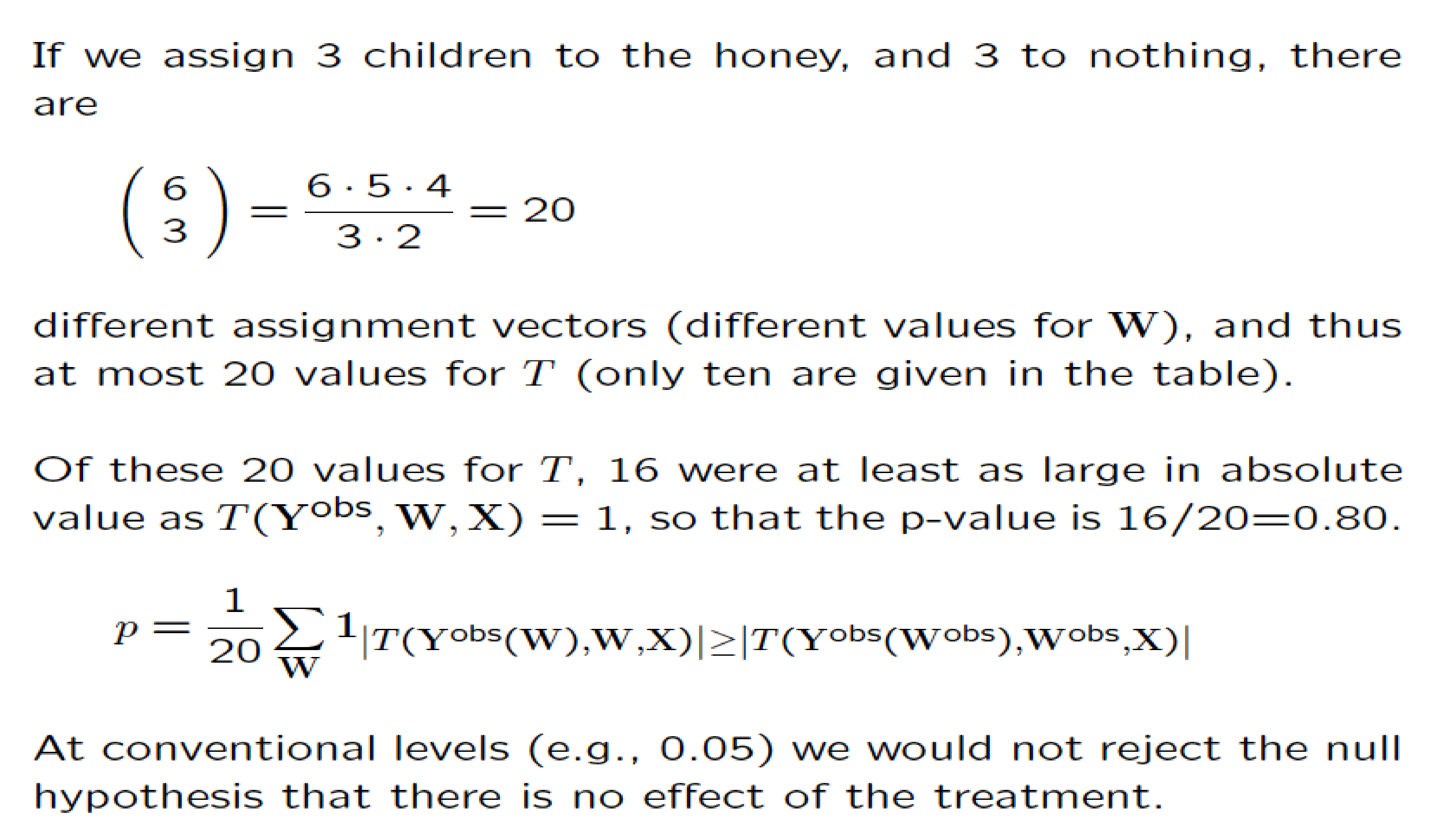
Filling in the "missing values"



Randomization Distribution for Two Statistics

						Statistic: Absolute Value of Difference in Average	
W_1	W_2	W_3	W_4	W_5	W_6	levels (Y_i)	ranks (R_i)
0	0	0	1	1	1	-1.00	-0.67
0	0	1	0	1	1	-3.67	-3.00
0	0	1	1	0	1	-1.00	-0.67
0	0	1	1	1	0	-1.67	-1.67
0	1	0	0	1	1	-0.33	0.00
0	1	0	1	0	1	2.33	2.33
0	1	0	1	1	0	1.67	1.33
0	1	1	0	0	1	-0.33	0.00
0	1	1	0	1	0	-1.00	-1.00
0	1	1	1	0	0	1.67	1.33
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Filling in the “missing values”



If we assign 3 children to the honey, and 3 to nothing, there are

$$\binom{6}{3} = \frac{6 \cdot 5 \cdot 4}{3 \cdot 2} = 20$$

different assignment vectors (different values for $\mathbf{W}$), and thus at most 20 values for T (only ten are given in the table).

Of these 20 values for T, 16 were at least as large in absolute value as $T(\mathbf{Y}^{\text{obs}}, \mathbf{W}, \mathbf{X}) = 1$, so that the p-value is 16/20=0.80.

$$p = \frac{1}{20}\sum_{\mathbf{W}} 1_{|T(\mathbf{Y}^{\text{obs}}(\mathbf{W}),\mathbf{W},\mathbf{X})| \geq |T(\mathbf{Y}^{\text{obs}}(\mathbf{W}^{\text{obs}}),\mathbf{W}^{\text{obs}},\mathbf{X})|}$$

At conventional levels (e.g., 0.05) we would not reject the null hypothesis that there is no effect of the treatment.

Assignment Mechanism

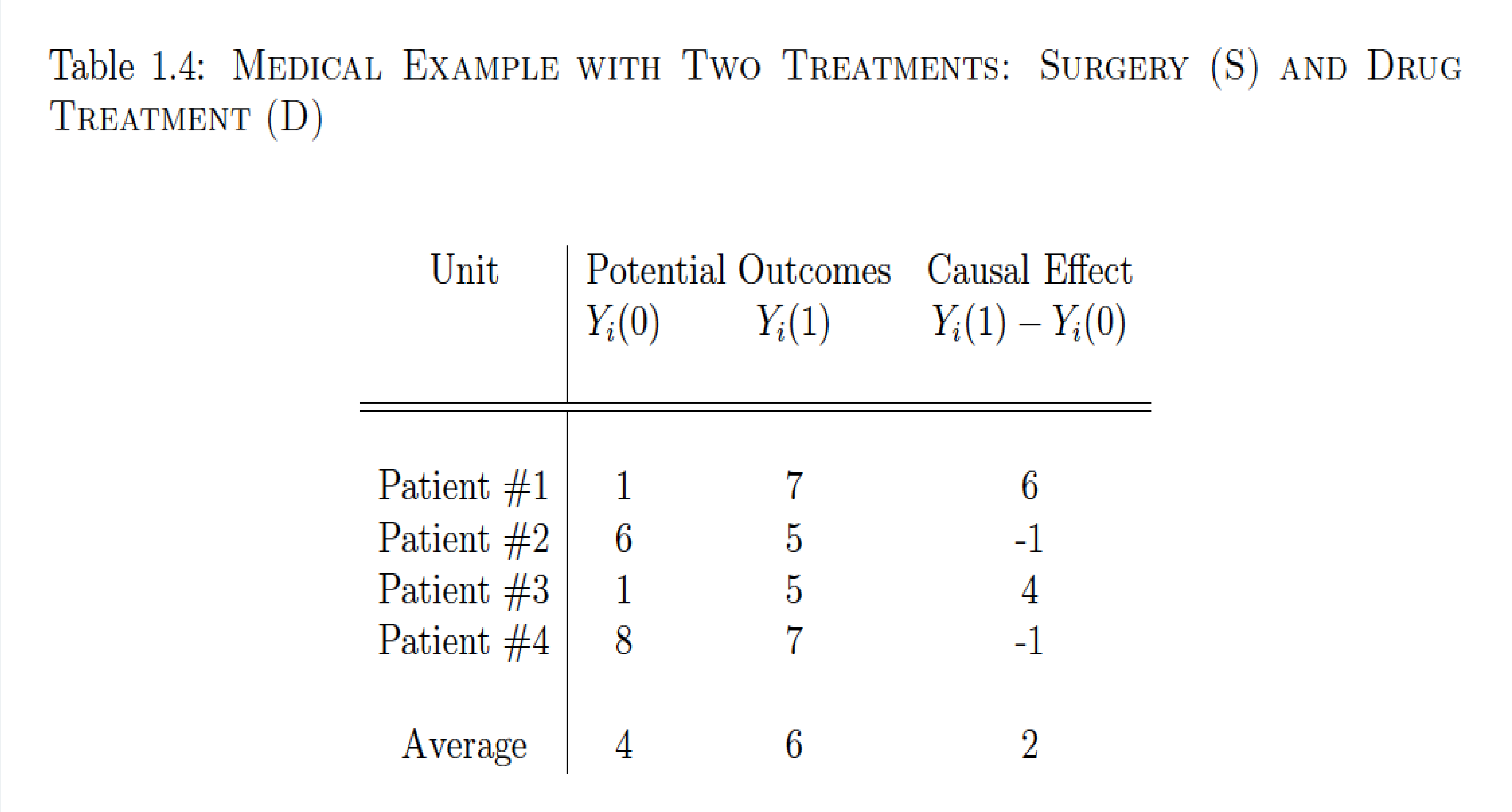


Table 1.4: MEDICAL EXAMPLE WITH TWO TREATMENTS: SURGERY (S) AND DRUG TREATMENT (D)

Unit	Potential Outcomes		Causal Effect
	$Y_i(0)$	$Y_i(1)$	$Y_i(1) - Y_i(0)$
Patient #1	1	7	6
Patient #2	6	5	-1
Patient #3	1	5	4
Patient #4	8	7	-1
Average	4	6	2

Assignment Mechanism

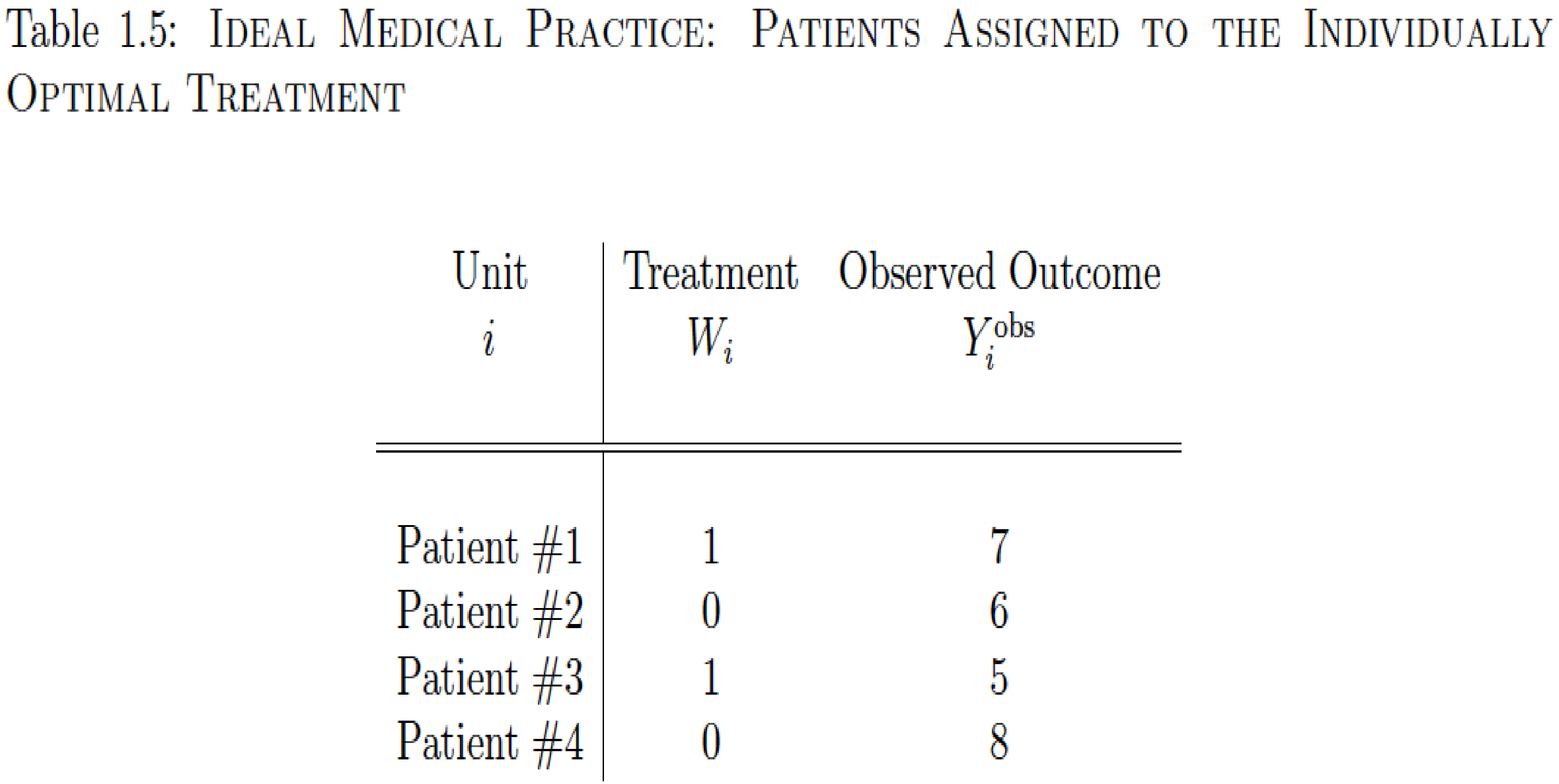


Table 1.5: Ideal Medical Practice: Patients Assigned to the Individually Optimal Treatment

Unit i	Treatment W_i	Observed Outcome Y_i^{obs}
Patient #1	1	7
Patient #2	0	6
Patient #3	1	5
Patient #4	0	8

Assignment Mechanism

- How is it determined that which units get treatment or, equivalently, which potential outcomes are realized (and which ones are missing)?

- Assignment mechanism is so crucial that causal inference depends on the assumptions concerning this mechanism.

- So, we arrive at why we randomize. It turns out that a “classical randomized experiment” provides the only assignment mechanism that satisfies the desired criteria for the estimation of causal effects “by design”.

Assignment Mechanism

A key issue for identifying and estimating causal effects is the assignment mechanism. Why do some units receive one treatment and others a different treatment? Is this random? Is this related to the outcomes given treatment or not? Who makes the decision?

The starting point in biostatistics is the randomized experiment, where assignment is completely random.

Social scientists often worry about selection bias. Units (individuals) who receive the treatment do so because they are different from individuals who do not receive the treatment. A classic example is schooling. Individuals who choose to go to college are different from individuals who choose not to go to college. They may have had higher earnings irrespective of their education levels. Randomization rules out selection bias.

Restrictions on the Assignment Mechanism

- We want three main restrictions on the assignment mechanism in order to estimate causal effects:

1. ***Individualistic*** assignment
    1. My probability to be assigned to treatment or control depends only on my pre-treatment covariates and my potential outcomes
2. ***Probabilistic*** assignment:
    1. Every unit has a positive probability to be assigned to treatment or control
3. ***Unconfounded*** assignment
    1. My assignment does not depend on my potential outcomes

Types of Studies

- A “classical randomized experiment” has (is) an assignment mechanism that satisfies all three restrictions listed above and the researcher knows and controls the functional form of the assignment mechanism.
  - Causal effects are straightforward to estimate and more often than not it is possible to do finite sample inference.

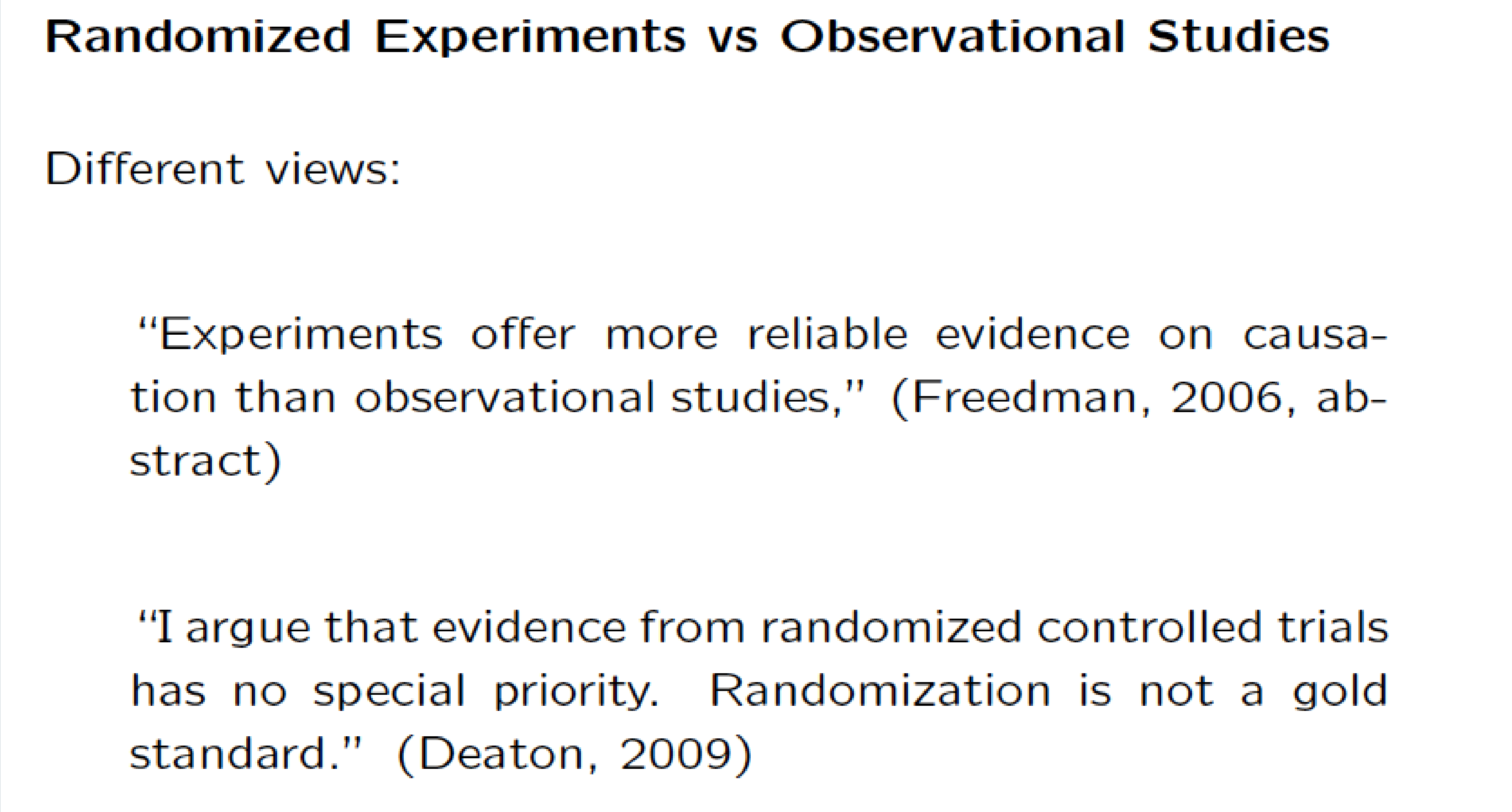
Types of Studies

- An assignment mechanism corresponds to an "observational" study if the assignment mechanism is unknown.
- A "regular assignment mechanism" is an observational study that satisfies all three restrictions, but by "assumption" rather than "design."
  - Need an initial "design" stage where covariate (pre-treatment) balance is assessed and sought.
  - In well-designed observational studies, you have many covariates that are associated with both assignment to treatment and potential outcomes.
  - Adjusting for these is sufficient to draw causal inferences…

Types of Studies

- An irregular assignment mechanism violates at least one of the three restrictions mentioned above. There are a number of interesting and tractable cases:
  - Non-compliance in randomized experiments necessitates the use of ***instrumental variables*** techniques, and, hence, invoking additional assumptions (such as exclusion restrictions)
  - Another interesting case is in circumstances where the probabilistic assignment is violated → ***regression discontinuity*** designs (as good as random assignment around the threshold)
  - Finally, we can think of using *pre* and *post* data for both the treatment and control group using panel (longitudinal) data: may be able to assume unconfoundedness given *pre* data → ***difference in differences*** method.

Common Critiques of Randomized Experiments



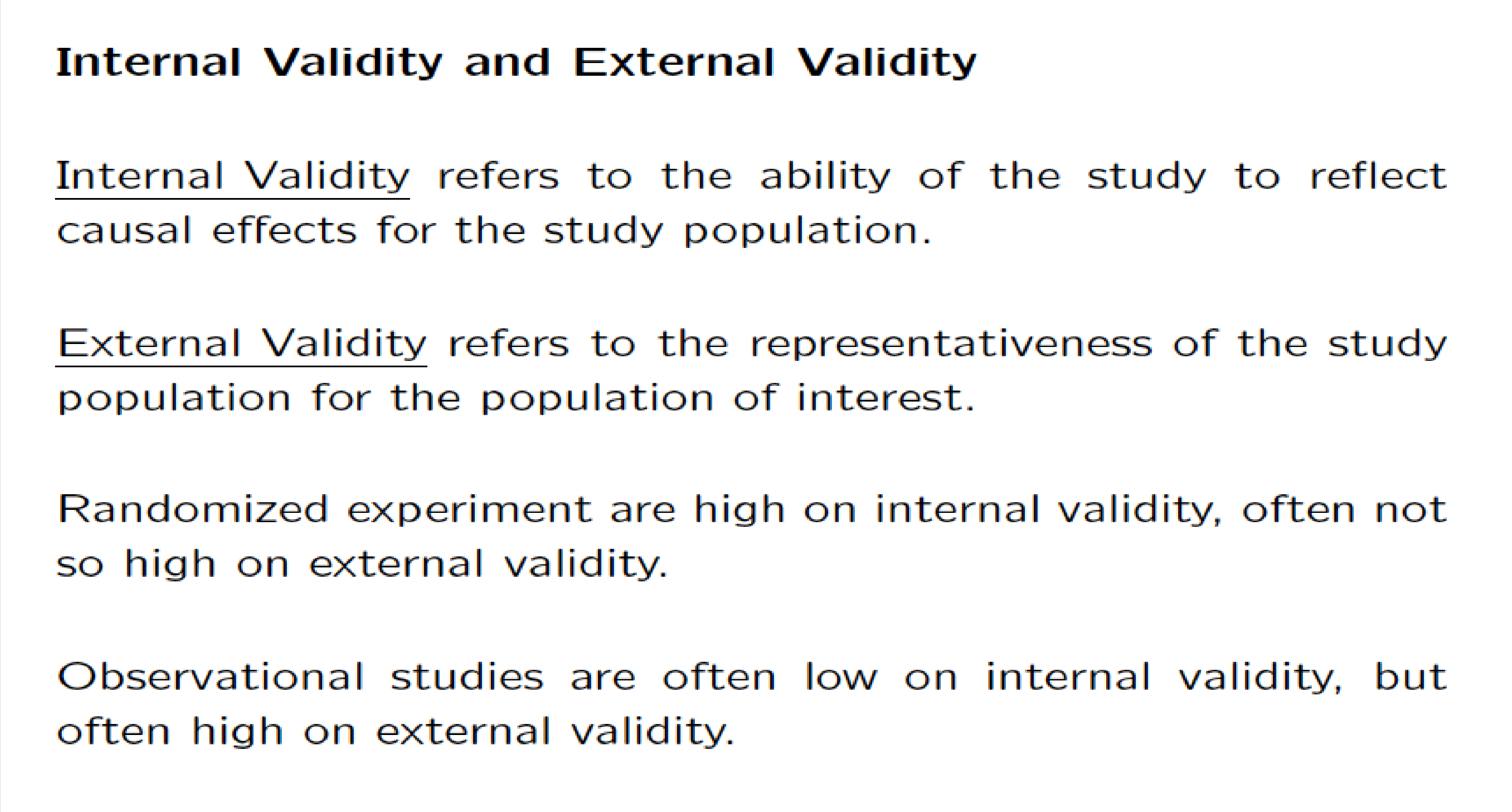
Randomized Experiments vs Observational Studies

Different views:

> "Experiments offer more reliable evidence on causation than observational studies," (Freedman, 2006, abstract)

> "I argue that evidence from randomized controlled trials has no special priority. Randomization is not a gold standard." (Deaton, 2009)

Common Critiques of Randomized Experiments



Internal Validity and External Validity

Internal Validity refers to the ability of the study to reflect causal effects for the study population.

External Validity refers to the representativeness of the study population for the population of interest.

Randomized experiment are high on internal validity, often not so high on external validity.

Observational studies are often low on internal validity, but often high on external validity.

Common Critiques of Randomized Experiments

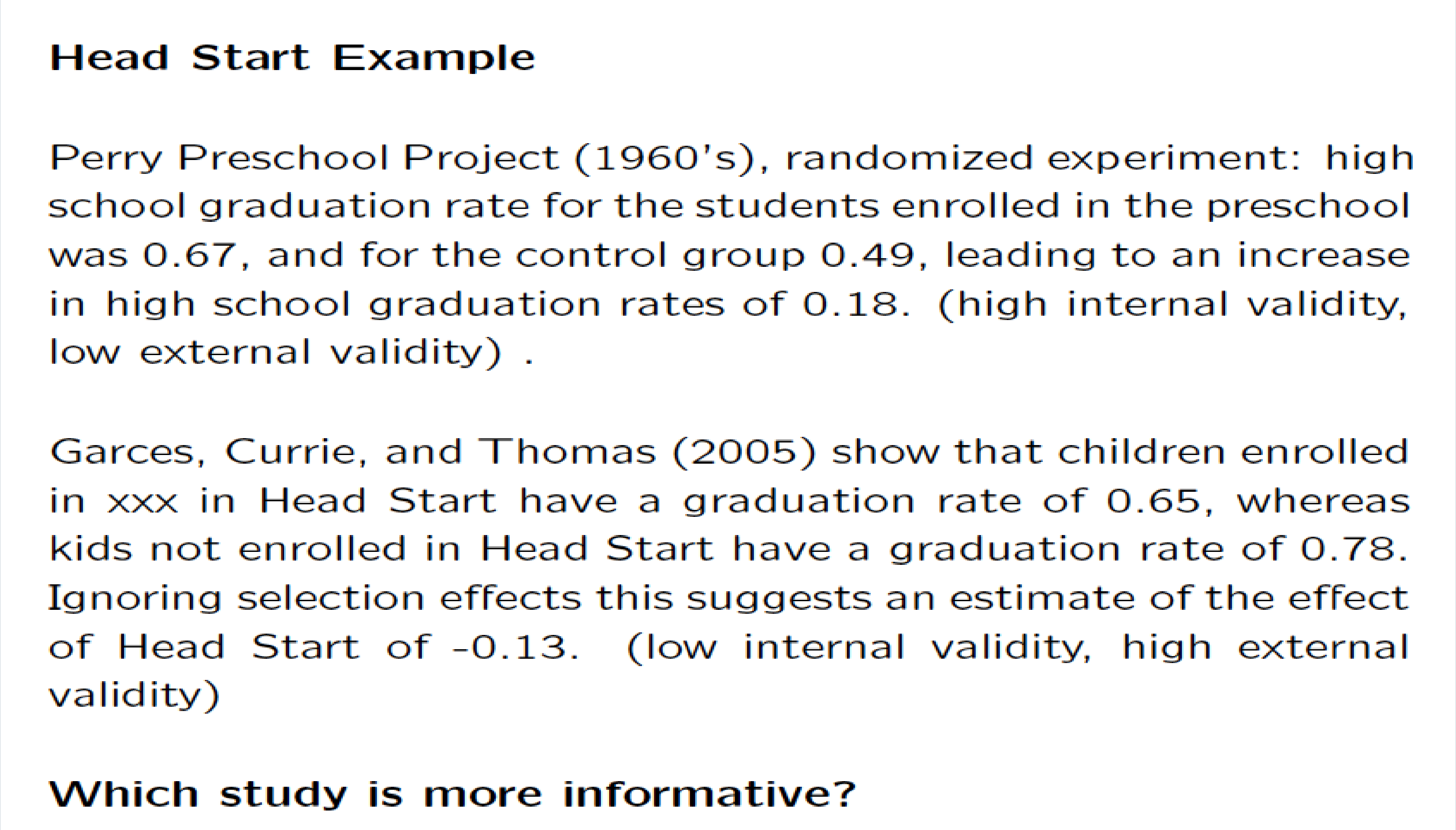
Campbell and Stanley (1963) claims

> "that studies should be judged primarily by their internal validity and only secondarily by their external validity."

Manski (2012) writes that

> "from the perspective of policy choice, it makes no sense to value one type of validity above the other."

Common Critiques of Randomized Experiments



Head Start Example

Perry Preschool Project (1960's), randomized experiment: high school graduation rate for the students enrolled in the preschool was 0.67, and for the control group 0.49, leading to an increase in high school graduation rates of 0.18. (high internal validity, low external validity) .

Garces, Currie, and Thomas (2005) show that children enrolled in xxx in Head Start have a graduation rate of 0.65, whereas kids not enrolled in Head Start have a graduation rate of 0.78. Ignoring selection effects this suggests an estimate of the effect of Head Start of -0.13. (low internal validity, high external validity)

Which study is more informative?

Common Critiques of Randomized Experiments

- ***Ethical issues*** – the importance of equipoise…
- ***Hawthorne and John Henry effects*** – these can be strong presences in certain contexts:
  - Hand-washing
  - School management/teacher incentive interventions…
- **Other?**

Part II

TYPES OF RANDOMIZED EXPERIMENTS

Opportunities to Randomize

In order to evaluate the impact of a program or policy, our randomly selected treatment group must have more exposure to the program than the comparison group. We can choose:

1. **Access:** Which people will be offered access to a program (***lottery***; lottery around a cut-off)
2. **Timing of access:** When to provide access to the program (***phase-in*** design)
3. **Encouragement:** Who will be given encouragement to participate in the program (***encouragement*** design)

When is it possible to randomize?

1. **New program design:** identified a problem; test alternative solutions.
2. **New program:** don't know impacts; random allocation fair
3. **New services:** randomize who has access
4. **New people/locations:** extend cut-off or location
5. **Oversubscription:** demand greater than supply
6. **Undersubscription:** offer encouragement
7. **Rotation:** can only cover a certain proportion at a given time
8. **Admission cut-offs:** often arbitrary
9. **Admission in phases:** resources will grow over time

Opportunity to randomize	**Example**
Program design	NGO wants to tackle obesity but not sure what the program should look like
New service	Insurance company want to introduce new insurance product for farmers
New people	Oregon has money to add more people to its medicaid program
New location	Microcredit company wants to expand to a new location
Oversubscription	More families sign up for education vouchers than government can fund

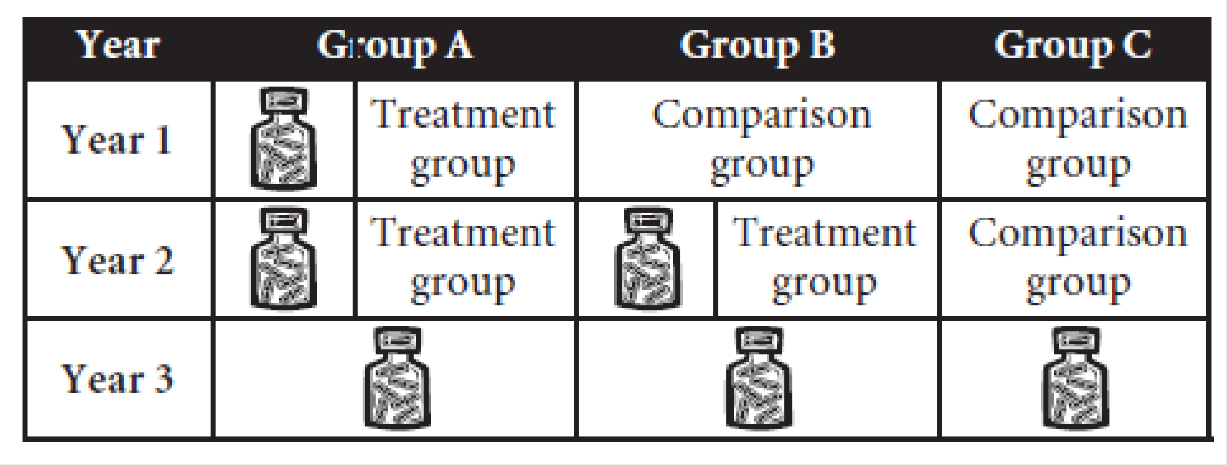
Opportunity to randomize	**Example**
Undersubscription	Lottery for conscription during Vietnam War
Rotation	Communities take turns to host an event
Admission cut off	Scholarship has merit cut off and ability to randomize admission just below and above
Admission in phases	A program builds 200 new schools but can only build 50 each year over a 4 year period

Before types of experiments, a word about ***ethics***

1. Just because we can randomly assign people to interventions, it does not mean we should. Consider:
    1. Equipoise
    2. Beneficence
    3. Privacy concerns
2. Resources:
    1. Development Impact blog: "Ethical issues with randomized experiments and other research,"
    2. Asiedu et al. (PNAS 2021): "A call for structured ethics appendices in social science papers."

Main study types

1. Lottery (classic A/B test; static experiment)
   1. Unconditional (Malawi CCT/UCT experiment)
   2. Around a cutoff or within a band (Nigeria Business Plan Competition)
2. Phase-in design (Worms, ECMA 2004)



Year	Group A		Group B		Group C
Year 1	💊	Treatment group	Comparison group		Comparison group
Year 2	💊	Treatment group	💊	Treatment group	Comparison group
Year 3	💊		💊		💊

Main study types

3. Encouragement design (Impacts of Econ Blogs)

- Useful when the ITT effects is not the main estimand of interest, but the LATE is.

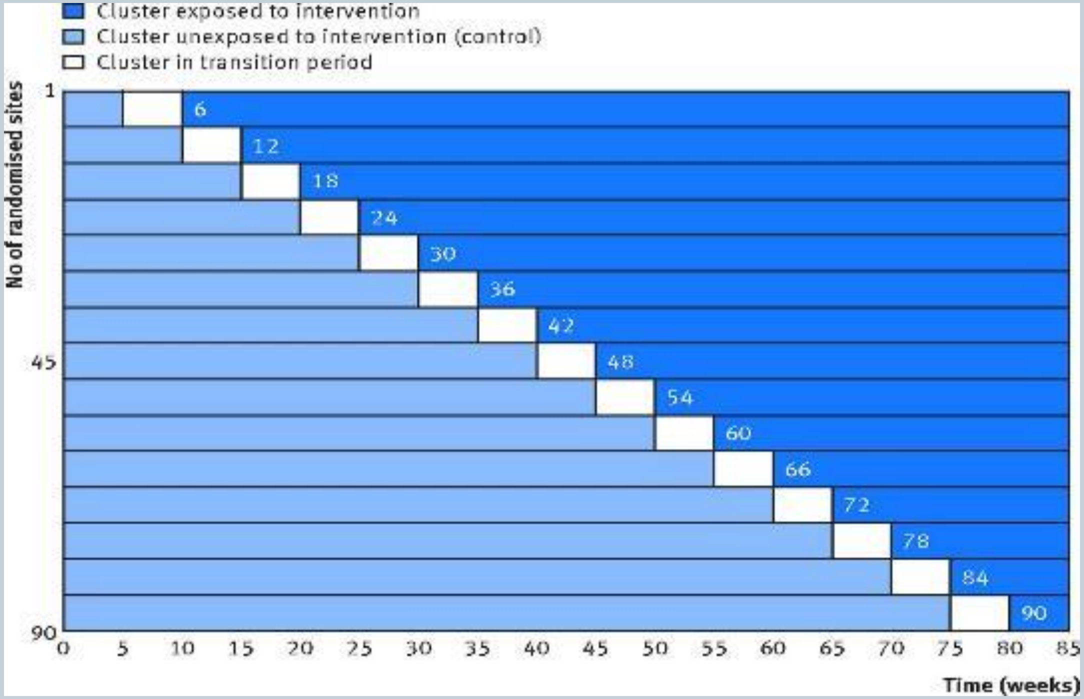
- Creates differential exposure when access to intervention is open to everyone (can't have a pure control group)

  - Randomized encouragement needs to satisfy IV needs

  - Can only estimate the LATE

Other (interesting) study types

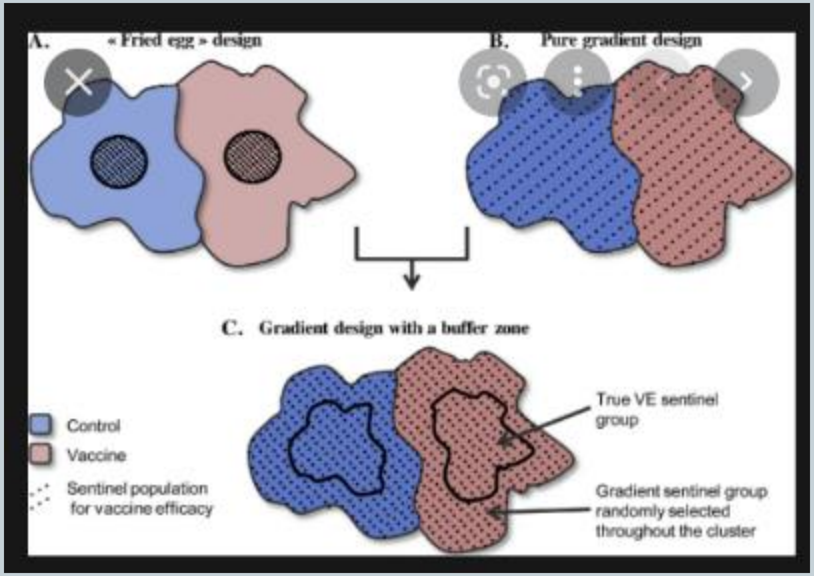
4. **Step-wedge design (Gambia Hep B vaccine)**
   - Similar to phase-in, more common in biomedical trials
   - Have to worry about time trends and anticipation effects



Other (interesting) study types

5. Fried-Egg design

- A version of the cluster-randomized (controlled) trials, where the whole “fried egg” assigned to treatment (or control) but only the yolk (in the center) is used for assessing impacts (free of spillovers from neighboring clusters).


A.
« Fried egg » design
B.
Pure gradient design
C. Gradient design with a buffer zone
Control
Vaccine
Sentinel population for vaccine efficacy
True VE sentinel group
Gradient sentinel group randomly selected throughout the cluster

Clustered vs. Individual Randomization (***SUTVA***)

- Remember the aspirin example from earlier:
  - Does my headache depend on you taking an aspirin? Probably not, but maybe…
- SUTVA rules out interference!
- Many cases where SUTVA is violated:
  - Pandemic
  - Guard rows in agricultural experiments
  - GE effects (job training or policing having a real effect or a displacement one?)
  - Peer effects in classrooms or schools

Clustered vs. Individual Randomization (***SUTVA***)

- Such effects can completely invalidate findings from RCTs that have not taken the possibility of interference into account during study design…

- In such cases, we take measures to minimize interference (or spillovers)

- Main method is to design a clustered-RCT.
  - Remember that there is no interference between clusters is still an assumption (although you can test this, ex post, with sufficient random variation in "distance" between clusters.

Clustered vs. Individual Randomization (***SUTVA***)

- From an implementation perspective, c-RCTs can be easier to pull off than individual randomization of treatment assignments:
  - e.g., jealousy within social networks, John Henry effects, etc.

- However, statistical power is significantly lower in c-RCTs (if intra-cluster correlation is high)

Factorial Designs

- Sometimes we want to know how two interventions interact with each other.
  - These are different than "packaged interventions," created to cause maximum impact on the outcome.
  - Also, different than trials with multiple intervention arms.
- Are mentoring interventions for adolescent girls more effective if combined with small cash transfers?
- Are ultra-poor programs more effective with an additional component (CBT, ongoing support, etc.)?

Factorial Designs

- In such cases, we run factorial (2x2, or n1Xn2) experiments…

- ***But, you have to design and analyze them correctly!***

Factorial Designs

- *$Y = a + b1*T1 + b2*T2 + b3*T1*T2 + e$ (long model)*

- *$Y = k + k1*T1 + k2*T2 + e1$ (short model)*

- ***You must design the evaluation, so that you can test b3=b1 with power.***

- ***Often, the short model does not describe a world we are interested in …***

- ***If you run the short model, the false rejection rates are high – even for modest interaction effects (|b3|>0)***

- ***Bonus general point: depending on the power of your study, p-value=0.17 can be evidence against the null (low power) OR p-value=0.04 can be evidence for the null.***

Factorial Designs

- Remember: you can always leave the interaction cell empty (*caveat: $N_c > N_1 = N_2$*)

- You can also choose to NOT have a control group (*more unconventional*)

- References:
  - (What) Should you do (with) experiments with factorial designs?
  - Muralidharan, Romero, and Wütrich (2022)
  - Be careful with inference from 2x2 experiments and other cross-cutting designs
  - Why p-values should be interpreted as p-values and not as measures of evidence

Part III

ADAPTIVE EXPERIMENTS

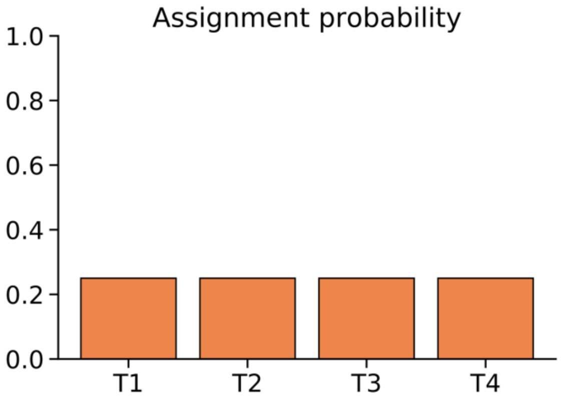
Adaptive Experiments

- So far, we have discussed static RCT designs. In contrast, an adaptive design may, based on interim analysis of the trial's result, change the allocation of subjects to treatment arms.
- They require, however, the measurement of the outcome (or a very good proxy of it) within a short period, so that assignment probabilities can be adapted.
- More suited to contexts, in which there is a rolling (continuous) enrollment of subjects into the trial…

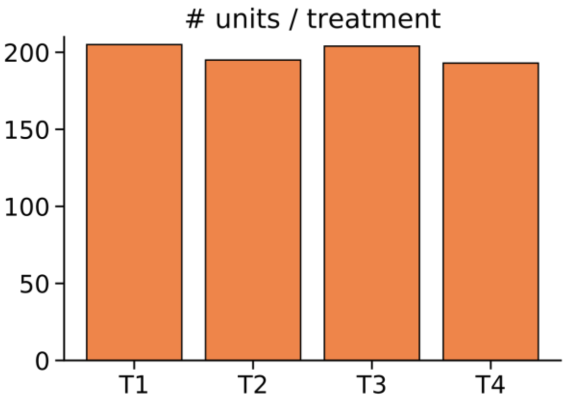
Adaptive Experiments

- They can help with:
  - Making the experiment more efficient (faster, smaller sample size – but not always)
  - Minimize “regret”: people in the study are more likely to be assigned to treatment arms beneficial for them
  - Generate more precise estimates for the “winner”
  - Adapt treatment assignments “contextually.”

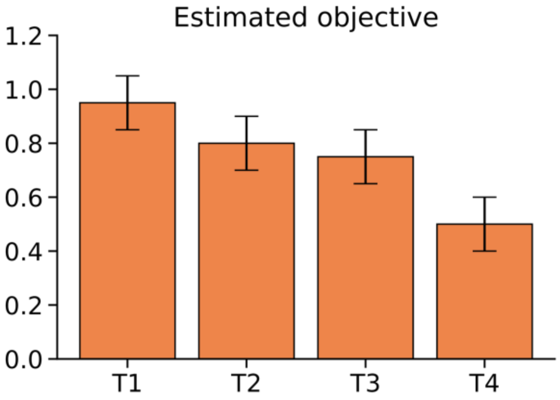
Randomized control trials



Fixed probability of assignment to each treatment*.



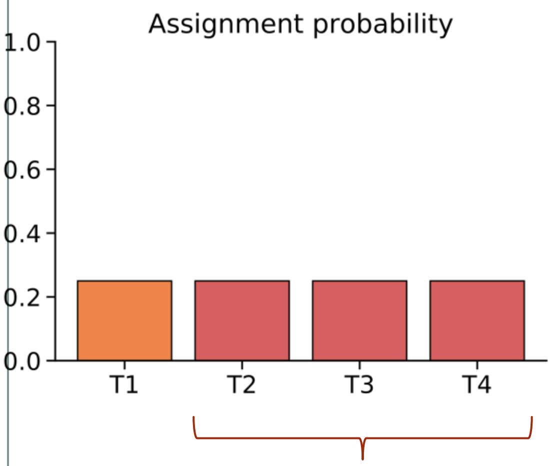
Roughly equal number of units assigned to each treatment.

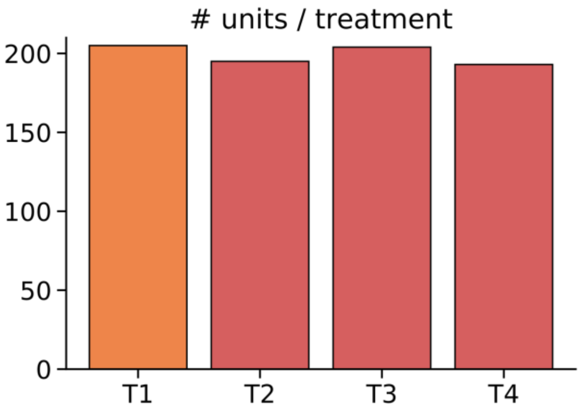


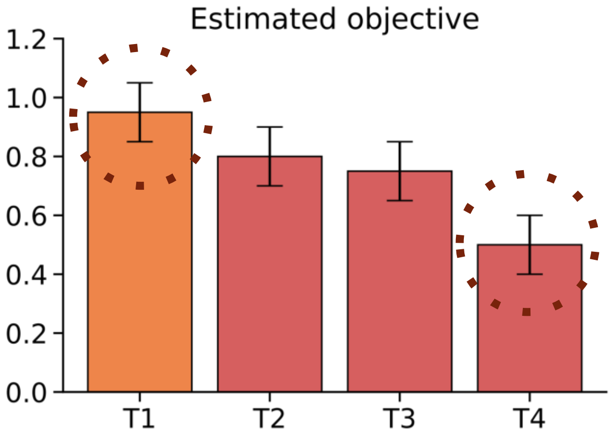
Treatment value estimate

*Note: for illustration only.

Randomized control trials





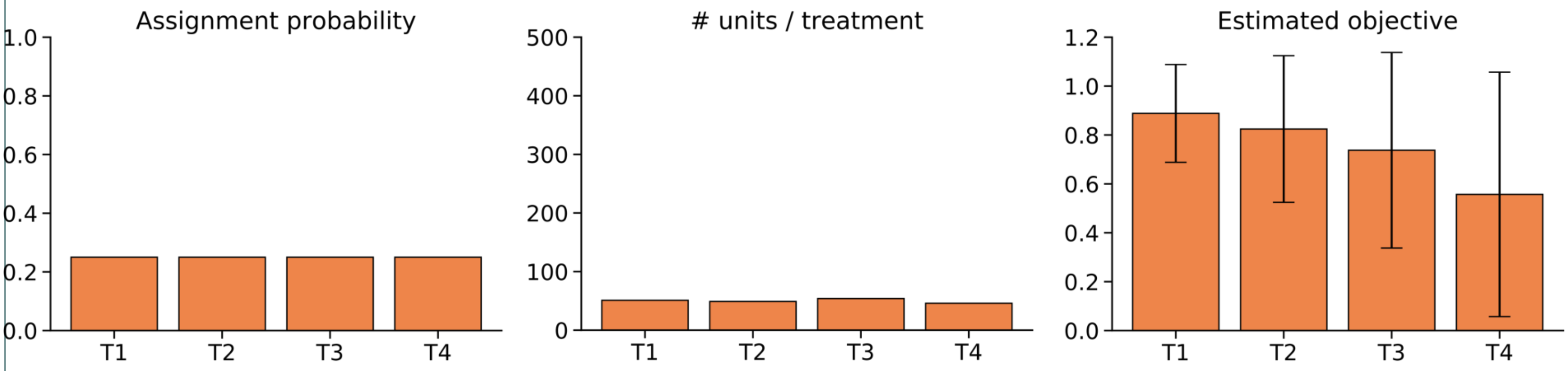


Many individuals assigned to suboptimal treatments (regret).

Good treatments not necessarily estimated more accurately than bad ones.

Adaptive experiments

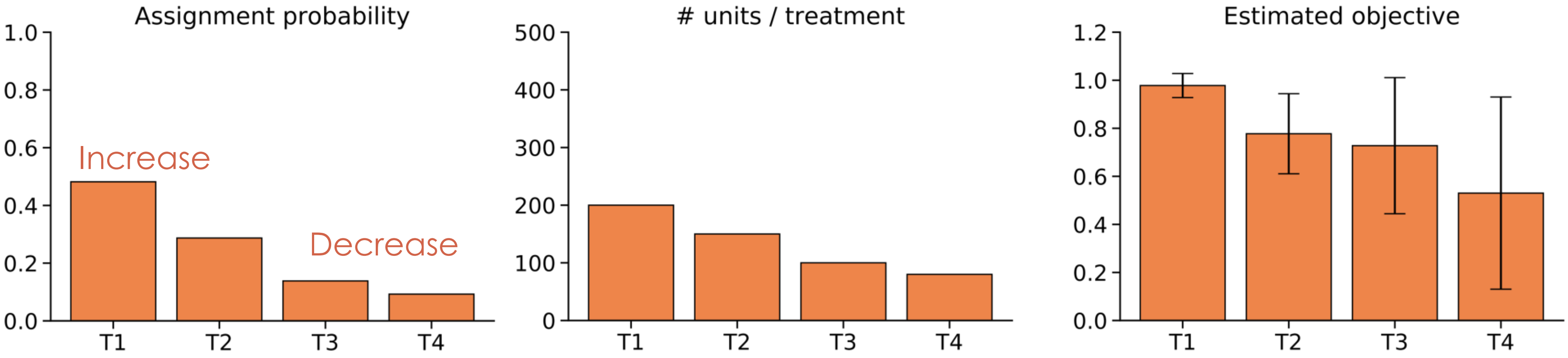
(example: multi-armed bandits)



Step 1: At the *beginning* of the experiment, assign treatments uniformly at random

Adaptive experiments

(example: multi-armed bandits)



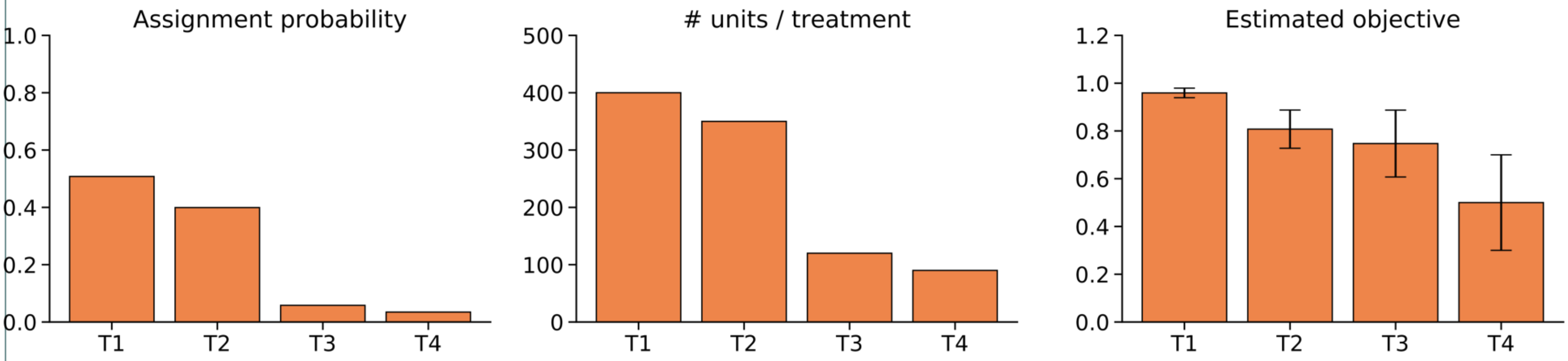
Increase

Decrease

Step 2: Once some data has been collected, increase the probability of assignment to more promising arms.

Adaptive experiments

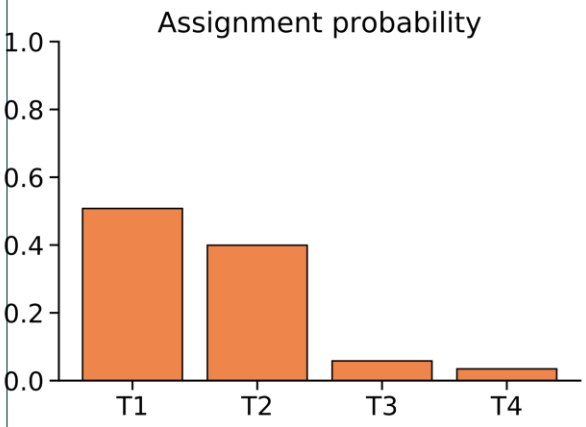
(example: multi-armed bandits)



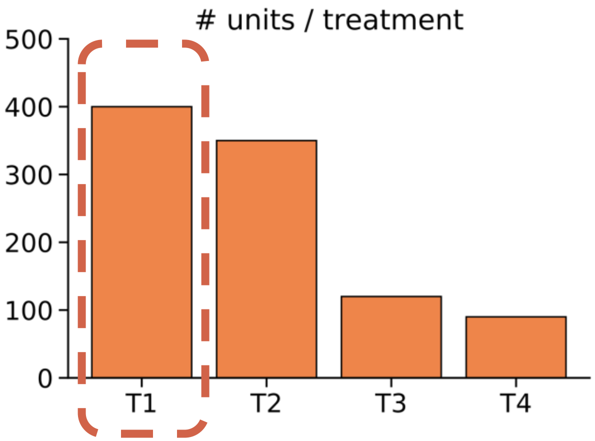
Step k: Repeat this procedure in batches, increasing probabilities of assignment as we become more certain about which treatments are good.

Adaptive experiments

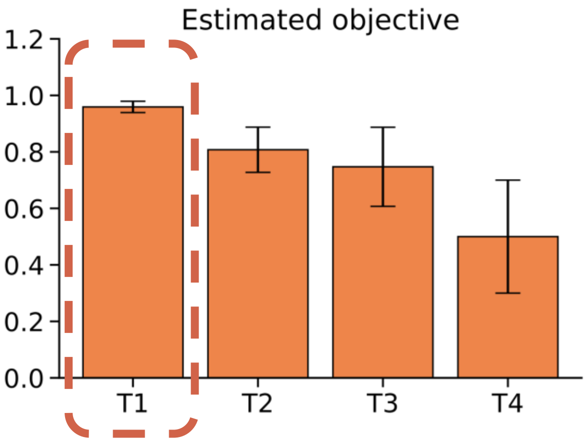
(example: multi-armed bandits)



As experiment progresses, suboptimal treatments are assigned less frequently…



…in the end, more observations assigned to optimal treatments (lower regret).

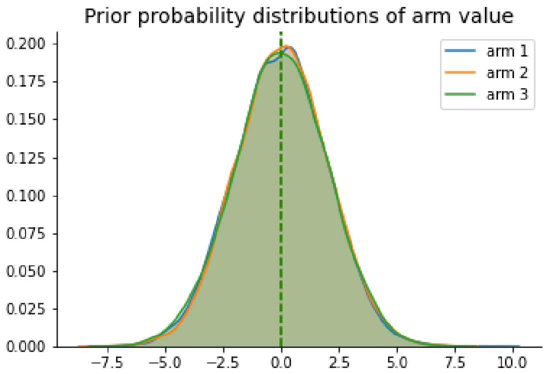


Tighter confidence intervals around optimal treatment value estimates.

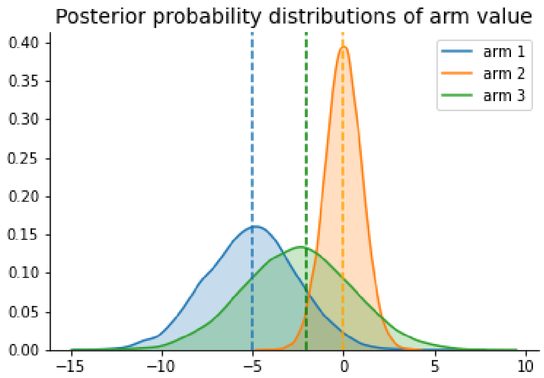
Computing assignment probabilities

Useful heuristic: **Thompson Sampling**

1. Start with a **prior** distribution on arm values.
2. Collect first batch of data by assigning treatments uniformly at random.
3. Observe outcomes and **update the posterior** distribution of arm values.
4. Next batch, assign treatments according to their **posterior probability of being optimal**.



P(arm 1 is optimal) = ⅓
P(arm 2 is optimal) = ⅓
P(arm 3 is optimal) = ⅓



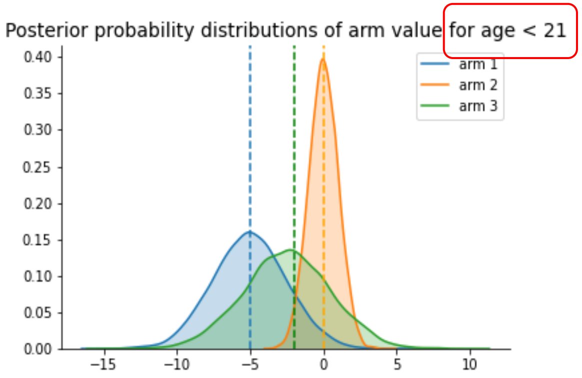
P(arm 1 is optimal | Data) = 0.05
P(arm 2 is optimal | Data) = 0.70
P(arm 3 is optimal | Data) = 0.25

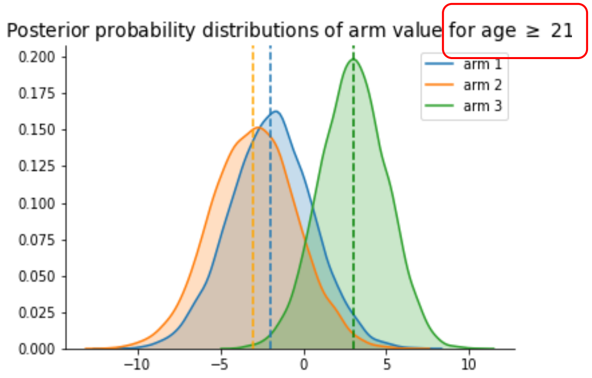
The **Thompson Sampling** heuristic dictates that these should be the assignment probabilities.

Adaptive experiments

(**contextual** bandits)

When personal characteristics (*contexts*) are observed, the assignment probabilities can be conditional on them.





As we gather more data, we are better able to **personalize** treatments.

Because bandit algorithms maximize the welfare of individuals in the experiment, they can be desirable from an ethical standpoint.

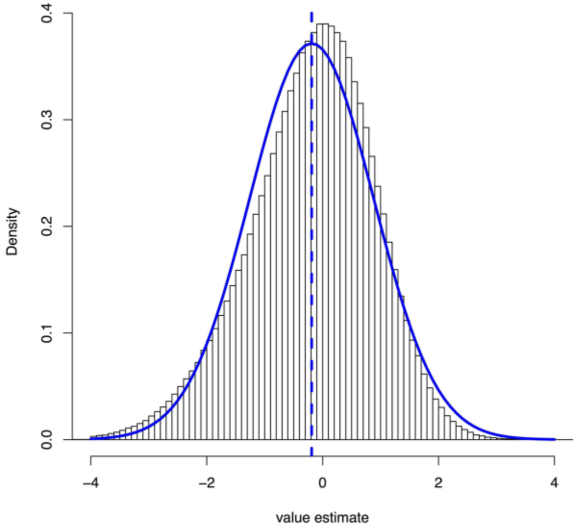
Caveats

In an adaptive experiment, collected data are **not independent**.

Usual methods for inference will often give the wrong answer. More sophisticated methods are needed.

This is an area of active research*

Luedtke and van der Laan (2016)
Deshpande, Mackey, Syrgkanis, Taddy (2017)
Hadad, Hirshberg, Zhan, Wager, **Athey** (2019)
Howard, Ramdas, McAuliffe, Sekhon (2019ab)
Zhang, Janson, Murphy (2020)

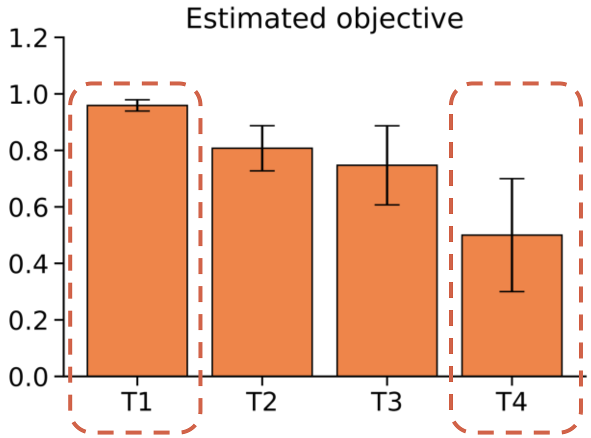


Example: distribution of the sample mean after an adaptive experiment. Estimates are biased and do not have a normal distribution.

Caveats

Need to make sure that the experimental design aligns with research objectives.

For example, if one goal is to be able to test the performance of the best treatment arm against a control, need to ensure that enough observations are collected from control.



Example: if one is expecting to test, e.g., whether T1 (the "best" arm) and T4 (the control arm) have the same mean, need to ensure enough observations in both T1 and T4.

Part IV

RECOMMENDED PRACTICES FOR ANALYSIS IN EXPERIMENTS

Baseline Balance and Attrition

- Perhaps, two of the most basic (and omnipresent) steps in the analysis of randomized experiments are demonstrating baseline balance and robustness for attrition (or participants lost to follow-up).

- Before we get to showing baseline balance, however, let's discuss what we can do at the design stage to address it.

Baseline Balance

- Matched pair randomization, matched quadruplets, and/or blocked (stratified) randomization?
- Perfect reference from David McKenzie on Development Impact last week…
- Main takeaways:

Baseline Balance

- Use one or two key variables with which you want to stratify your sample. You could have theoretical or practical guidance on treatment effect heterogeneity by, say, gender or location or other. E.g., Imbens recommends using cluster size to define strata.

- Within these form quadruplets, using either:
  - The outcome indicator (or and index of closely correlated outcome indicators), OR
  - A set of baseline covariates that predict the outcome well.

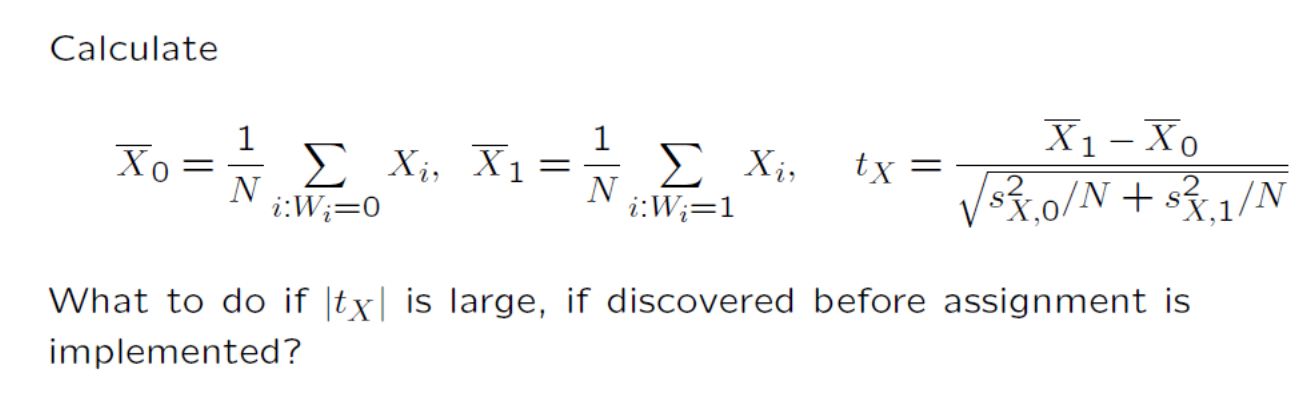
Baseline Balance

- If you do this, then you will have balance on the stratification variables, as well as the outcome variable (or an index of outcome variables).
    - If you have used Mahalonobis distance, you will also have balance on that score (and most likely all the variables that go into it).

- In this case, my recommendation is that it is OK to report baseline balance in Appendix Table A1.

Baseline Balance

- Should you re-randomize for baseline balance?



Calculate

$$\overline{X}_0 = \frac{1}{N}\sum_{i:W_i=0} X_i, \quad \overline{X}_1 = \frac{1}{N}\sum_{i:W_i=1} X_i, \qquad t_X = \frac{\overline{X}_1 - \overline{X}_0}{\sqrt{s^2_{X,0}/N + s^2_{X,1}/N}}$$

What to do if $|t_X|$ is large, if discovered before assignment is implemented?

- Two options:
  1. Randomize M times, implement assignment vector that minimizes the maximum t_{Xi}, $i \in 1, 2, \ldots, K$
  2. Re-randomize until all t-stats below a certain threshold.
- Pre-specify for randomization inference, and
- Don’t search over randomizations for best value.

Baseline Balance

- Again, if you do this, i.e., re-randomize, then you will have balance on the pre-specified variables.
  - In this case, my recommendation is that it is OK to report baseline balance in Appendix Table A1.

- If you have not done either of these strategies and either:
  - implemented unconditional randomization, OR
  - Did not implement the randomization yourself, then
- Report baseline balance in Table 1.

Baseline Balance

- A few things to remember:

1. If your study is well-powered, more likely to reject differences in covariate values between arms and vice versa: so, design well-powered studies, report MDEs, and have an idea of meaningful (vs. stat. sig.) differences.
2. Conduct tests of joint orthogonality of all variables (*F-tests for T & C; chi-squared tests using multinomial logit with multiple treatment arms*)

- You want high p-values in these tests, especially when your experiment is only modestly powered.

Attrition (or loss to follow-up)

- Attrition is the bane of RCTs!
  - In some ways, it is the biggest threat to clean identification in RCTs, as you cannot prevent it.
  - **The best way to deal with attrition is to minimize it! The smaller the number of observations lost to follow-up, the easier are the fixes…**
- Two issues with attrition:
  - Differential attrition ***in levels***, and
  - Differential attrition ***in baseline characteristics***.
  - If you can show evidence against both these, then you may be allowed to proceed with no corrections or sensitivity analysis.

Attrition (or loss to follow-up)

- How to present the evidence on attrition:
  - Use the same variables as presented in the baseline balance table (preferably pre-specified based on the outcomes, covariates that are prognostic of them, and sources of heterogeneity)…
  - Run a regression that is fully interacted (saturated) with treatment(s) and its(their) interaction with (centered/de-meaned) covariates.
  - Report F-tests of joint orthogonality – separately for the covariates and their interactions
    - Covariates influencing attrition is not a problem – it is expected.
    - Interactions significantly influencing attrition is a problem.
      - So is any level differences in attrition…

Attrition (or loss to follow-up)

- See examples of attrition analysis here and here…

- What to do when there is differential attrition:
  - Inverse propensity weights (generally do not change findings)
  - Manski and Lee bounds (former generally yields bound that are too wide to be informative, while the latter comes with some assumptions. See this blog post for a detailed discussion)
  - Kling-Liebman bounds (my preferred route)
    - In the study of a UCT program for Syrian refugees, hyperlinked above, we made this part of the main analysis (attrition very high!)

Should we do covariate adjustments in RCTs?

- There have been several concerns with covariate adjustments in randomized experiments:
  - Famous statistician David Freedman worried that precision could be hurt by adjustments.
  - Adjustment can also open the door to fishing, i.e., ad hoc specification searching (or p-hacking).
    - Remember that difference-in-differences, which used to be the norm, as well as ANCOVA (broadly defined) are both available to researchers, in addition to the choice of (not prespecified) covariates (see this blog post for a more detailed discussion)

Should we do covariate adjustments in RCTs?

- The first problem can come from two sources:
  - Unbalanced assignment to treatment vs. control
  - Strong treatment-effect heterogeneity in the adjusted covariate

- Winston Lin, in two classic Development Impact posts (Parts I & II), has shown that, adjustment cannot hurt precision if you regress:
  - Y on T, X - xbar, and T * (X - xbar), where xbar is the mean covariate value for the entire sample.
  - Then the coefficient on T estimates the average treatment effect (ATE) for the entire sample.

Should we do covariate adjustments in RCTs?

- "The main purpose of allowing [adjusting] for covariates in a *randomized* trial is defensive: to make it clear that analysis has met its scientific obligations." *John Tukey*
  - The researchers, whenever possible, should pre-specify the covariates to be used for adjustment (*caveat on LASSO…*)
  - Doesn't matter if the adjusted or the unadjusted specification is the 'main analysis,' as long as both are reported.
  - Select a concise K-vector of adjustments, from the set of covariates that are strongly prognostic of the outcome at follow-up (lagged value of the outcome variable is natural).
    - K<<N

Concluding remarks

- Analyze the RCT you designed!
  - Block/stratum fixed effects
  - Randomization inference (upcoming lecture)
  - Careful analysis and transparent discussion of balance, attrition, implementation problems, fully-interacted covariate adjustments (and unadjusted estimates), etc.
  - ***Many other aspects we did not cover (randomization mechanics, field work details, experimental design issues, missing data, etc.)***