Section 2: Introduction to identification, potential outcomes, causal relationships and random treatment assignment under fixed margins

Yotam Shem-Tov

Fall 2015

DGPs, Models, and Identification

- A *data generating process* (DGP) is a complete specification of the stochastic process generating the observed data.
- Knowledge of the DGP allows one to compute the likelihood of any realization of the data but is conceptually distinct from the distribution of observed data since it provides a description of the mechanism (or structure) giving rise to this distribution.
- Example: The following two DGPs generate the same likelihood of the observed variables $\{Y_i,\ X_i\}_{i=1}^N$,

$$Y_i = \epsilon_i, \quad \text{where} \quad \epsilon_i \sim N(0,1)$$

$$Y_i = \eta_i + \zeta_i, \quad \text{where} \quad \begin{pmatrix} \eta_i \\ \zeta_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{1}{\sqrt{2}} & 0 \\ 0 & \frac{1}{\sqrt{2}} \end{pmatrix} \right)$$

$\Rightarrow$ Both data generating processes above are different, however they yield the same distribution of Y_i, and the same probabilities for the observed values of Y_i.

Identification

- Identification analysis is the study of which structures are consistent with the joint distribution of observed variables.
- Denote by Z_i a vector of observed variables.
- Let $F_Z(z)$ denote the distribution of governing observed variables, and $F_\theta(z)$ the distribution function implied by a particular structure θ.
  - The distribution function of the observed data is $F_Z(z)$, the sample analogue is $\hat{F}_Z(z)$ and $\hat{F}_Z(z) \to F_Z(z)$.
  - $F_\theta(z)$ is the distribution function implied by a structure θ.
  - If two structures θ' and θ'' yield the same distribution, $F_{\theta'}(z) = F_{\theta''}(z)$. As both structures yield the same distribution of Z we cannot detect whether the correct structure is θ or θ' using $F_Z(z)$.

The *identified set* of structures

$$\Omega(F_Z(\cdot), \Theta) = \{\theta \in \Theta F_\theta(\cdot) = F_Z(\cdot)\}$$

where $F_Z(\cdot)$ is the true distribution of the data, and Θ is the space of possible structures θ.

Identification

- The structure θ is point identified if $\Omega(F_Z(\cdot), \Theta)$ is a singleton (i.e., the set contain only one element).
- Two structures of the data θ' and θ'' are *observationally equivalent* if $F_{\theta'}(z) = F_{\theta''}(z)$ for all $z \in \mathbb{R}^k$, where k is the dimension of z.
- If two structures are *observationally equivalent*, then we cannot distinguish between them regardless of the sample size.

Definitions from Wikipedia

Definition

Parameter: A number or vector that indexes a family of distributions
Example: the rate parameter in a Poisson distribution, or the potential outcomes in our causal model.

Definition

Identifiability: Let P_θ be a family of distributions indexed by θ. A function of θ is identifiable if $f(\theta_1) \neq f(\theta_2)$ implies $P_{\theta_1} \neq P_{\theta_2}$ for all θ_1, θ_2.

Definition

Estimability: A function $f(\theta)$ is estimable if there exist an estimator of $f(\theta)$ that is unbiased.

Theorem

If $f(\theta)$ is estimable then $f(\theta)$ is identifiable

The other direction does not hold. Estimability implies Identifiability, but Identifiability does imply estimability.

Example: Let $0 < p < 1$ and x be binomial with $P_p(x = 1) = p$. The function $f(\theta) = \sqrt{p}$ is identifiable, however $\sqrt{p}$ is not estimable.
Let $g(x)$ be some estimator. Then,

$$\mathbb{E}_p[g(x)] = (1-p)g(0) + pg(1)$$

This is a linear function in p, however $\sqrt{p}$ is not a linear function of p. So, $\mathbb{E}_p[g(x)] \neq \sqrt{p}$. Show that p is identifiable:
$\sqrt{\bar{x}} \to \sqrt{\mathbb{E}[x]} = \sqrt{p}$, when $n \to \infty$ we know $\mathbb{E}[x]$ and therefore can *point identify* $\sqrt{p}$.

Identification example: Mixture of Normal distributions

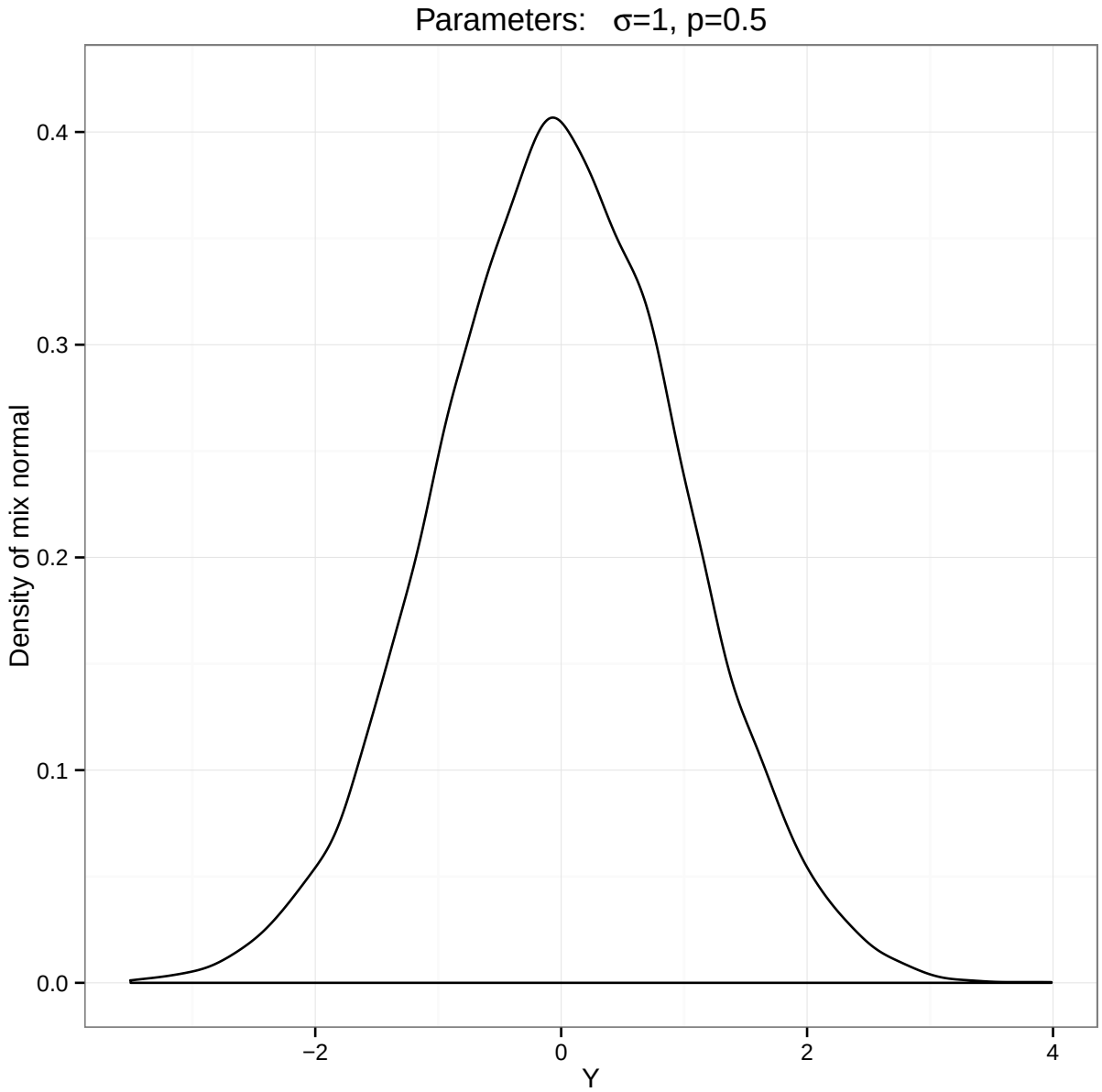
- Consider the following Model,

$$Y_i = \begin{cases} N(0,1), & \text{with probability } p \\ N(0,\sigma^2), & \text{with probability } 1-p \end{cases}$$

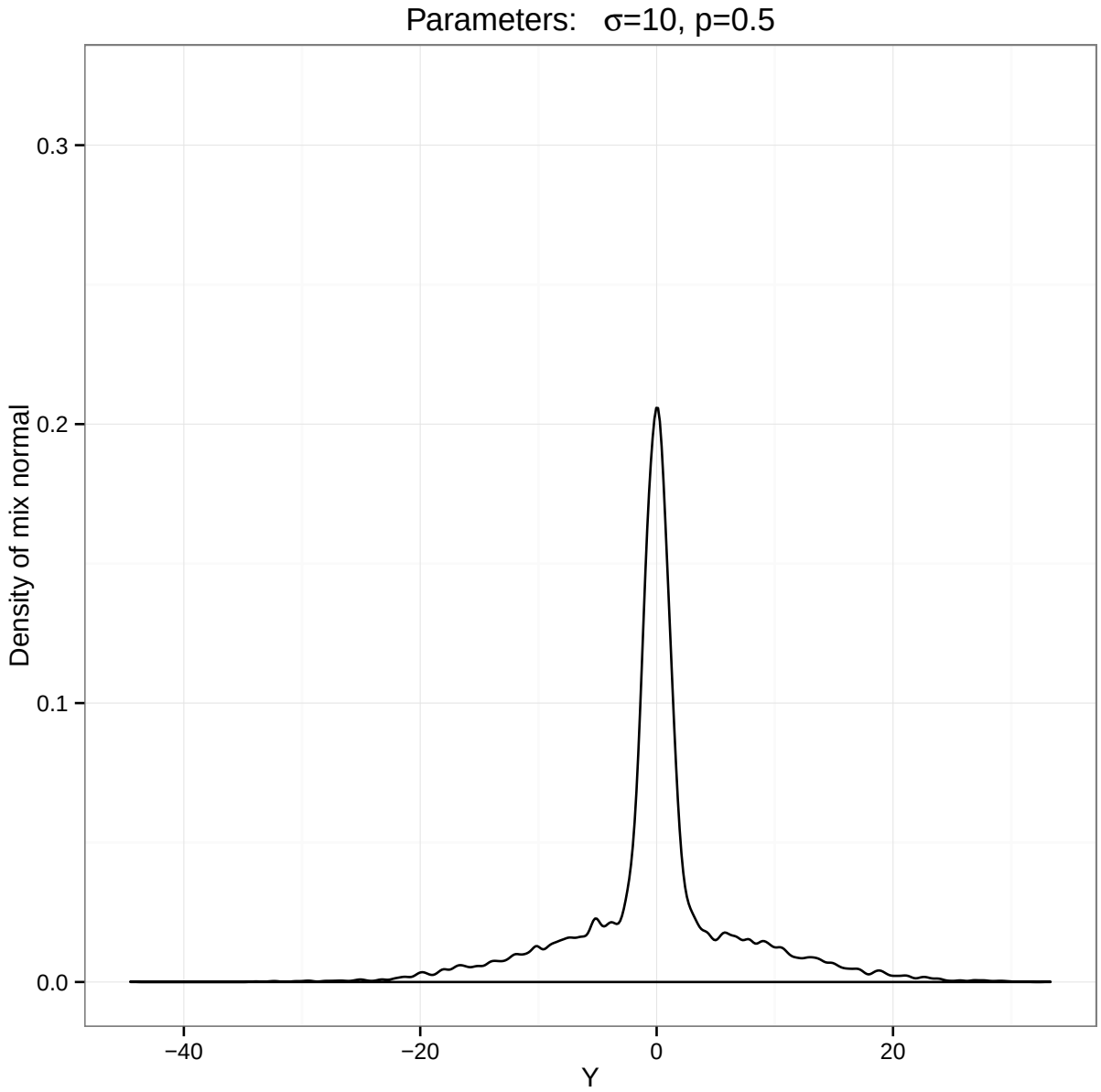
- This model can describe multiple DGPs, depending on different values of p and σ^2
- Describe a structure of the data, i.e., value of p and σ^2 such that at least one of the parameters is not identified.

When $\sigma^2 = 1$, the parameter p is not identified. Any value of $p \in [0,1]$ will yield the same distribution of Y. Therefore we cannot identify p in any possible data set, when $\sigma^2 = 1$.

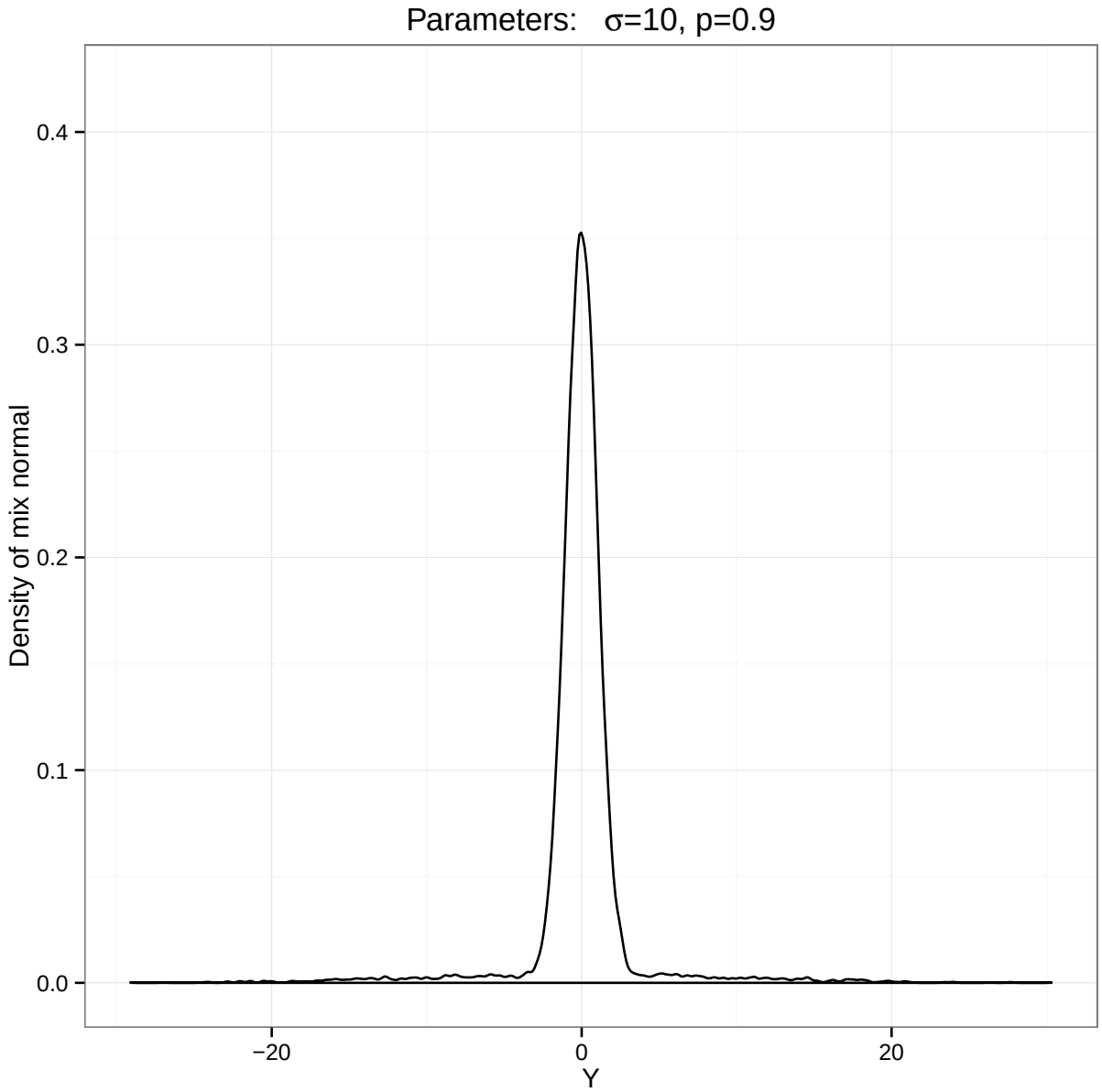
Identification example: Mixture of Normal distributions



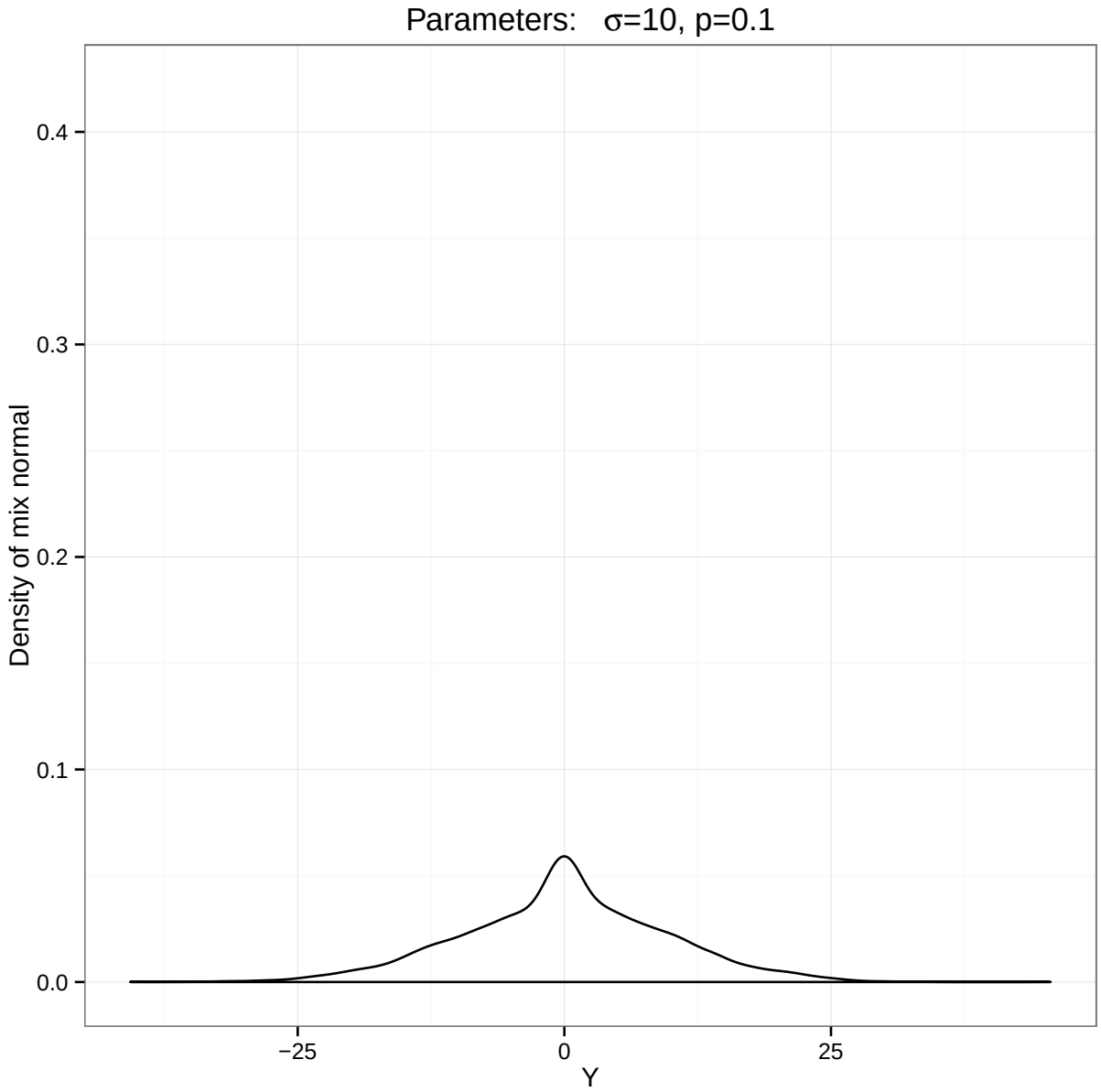
Identification example: Mixture of Normal distributions



Identification example: Mixture of Normal distributions



Identification example: Mixture of Normal distributions



R exercise: simulating data from a mixture of normals

Write code to simulate $N = 1000$ observations from the distribution below and plot the density using ggplot.

$$Y_i = \begin{cases} N(0,1), & \text{with probability } p = 0.5 \\ N(5,1), & \text{with probability } 1-p = 0.5 \end{cases}$$

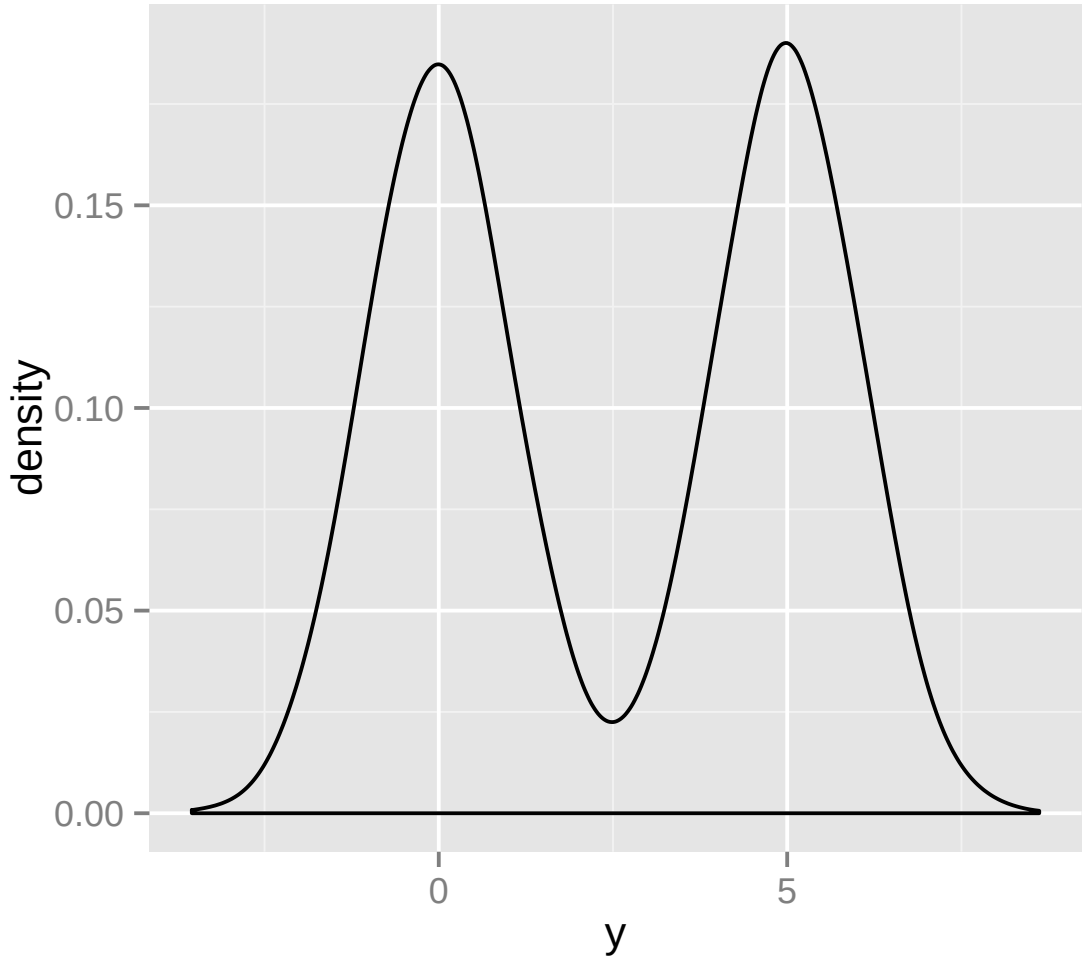
Hint: Load the package ggplot2, and use
ggplot()+geom_density(aes(x=data))

R exercise: Solution

```
require(ggplot2)
set.seed(12345)

n=10000
p=0.5
indicator = rbinom(n,size=1,prob=0.5)
normal0 <- rnorm(n,mean=0,sd=1)
normal1 <- rnorm(n,mean=5,sd=1)
y = normal0
y[indicator==1]=normal1[indicator==1]
ggplot()+geom_density(aes(x=y))
```

R exercise: Solution



This is refereed to by Jas as "The claw"

Definitions: Potential outcome

- Let T_i be an indicator variable whether individual i received treatment ($T_i = 1$) or control ($T_i = 0$)
- Let Y_{i1} be the potential outcome of individual i with treatment and Y_{i0} the potential outcome without treatment
- The observed outcomes are,

$$Y_i = T_i Y_{i1} + (1 - T_i) Y_{i0}$$

Group	Y_{i1}	Y_{i0}
$T = 1$	Observable: $Y_{i1} \vert T = 1$	Counterfactual: $Y_{i0} \vert T = 1$
$T = 0$	Counterfactual: $Y_{i1} \vert T = 0$	Observable: $Y_{i0} \vert T = 0$

- Throughout this slides (and future sections) I will use the following two notations for treatment and control interchangeably: $Y_i(1) \Leftrightarrow Y_{i1}$ and $Y_i(0) \Leftrightarrow Y_{i0}$.

Identification of causal relationships

- Causality can be thought of as the study of the joint distribution of potential outcomes $(Y(1), Y(0)$.
- Holland 1986, "No causation without manipulation"
  $\Rightarrow$ We cannot identify causal parameters without exogenous manipulation in the assignment of treatment.
- It is important to ask which structures (i.e., parameters) are of interest, and whether the manipulation of treatment allows to identify this objects of interest.

Definitions: Treatment effects

- I follow Imbens (2004) in the definitions of average treatment effects.
- Imbens (2004) is an excellent reading for the identification assumptions in regression and matching methods. This is highly recommended reading!
- The treatment effect on individual i is,

$$\tau_i = Y_{i1} - Y_{i0}$$

- There can be many parameters of interest. A few common parameters are, the sample average treatment effect, and the population average treatment effect.
- The difference is whether we are interested in the average treatment effect in this specific sample (conditional on the sample), or whether we want the average treatment effect in the population from which the sample we observe was drawn.

Definitions: Treatment effects

- The sample ATE, ATT and ATC are,

$$SATE = \sum_{i=1}^{N} (Y_{i1} - Y_{i0}), \quad SATT = \sum_{i=1}^{N} (Y_{i1} - Y_{i0} | T_i = 1)$$

$$SATC = \sum_{i=1}^{N} (Y_{i1} - Y_{i0} | T_i = 0)$$

- The population ATE, ATT and ATC are,

$$PATE = \mathbb{E}(Y_1 - Y_0), \quad PATT = \mathbb{E}(Y_1 - Y_0 | T_i = 1)$$

$$PATC = \mathbb{E}(Y_1 - Y_0 | T_i = 0)$$

- We can also be interested in the treatment effect conditional on a certain value of Y_0, for example:

$$\mathbb{E}(Y_1 - Y_0 | Y_0 \leq K)$$

- For example instrumental variables identify a local average treatment effect, i.e., the treatment effect for the population of compliers.

An identification example of average treatment effects

- Consider following DGP:

$$Y_i = p_i + D_i \cdot \tau \quad \Pr(D_i = 1) = \mathbb{I}\{p_i \leq k\}, \quad p_i \sim Unif[0,1]$$

  the researcher observes (Y_i, D_i), and cannot observe p_i.
- What is the distribution of Y_i when $D_i = 1$ and when $D_i = 0$?
  $Y_i|D_i = 1 = \tau + p_i|D_i = 1 \sim Unif[0,k]$
  $\Rightarrow Y_i|D_i = 1 \sim Unif[\tau, k+\tau]$,
  and $Y_i|D_i = 0 = p_i|D_i = 0 \sim Unif[k,1]$
- What is the PATT? is it equal to the PATE? Yes,
  $\mathbb{E}[Y_i(1) - Y_i(0)|D_i = 1] = \mathbb{E}[p_i + \tau - p_i|p_i \leq k] = \tau$
- Can we identify k from the data? Yes, $\bar{D} \to \Pr(D_i = 1) = k$.
- Can we identify τ? Is τ an estimable parameter? Yes

$$\mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0] = \mathbb{E}[p_i + \tau|p_i \leq k] - \mathbb{E}[p_i + \tau|p_i \geq k]$$
$$= \tau + \frac{k}{2} - \frac{1+k}{2} = \tau - \frac{1}{2}$$

  $\Rightarrow \bar{Y}_{D=1} - \bar{Y}_{D=0} + 1/2$, is an unbiased estimator for τ.

Median treatment effect

- Is the median treatment effect, $median(Y_{i1} - Y_{i0})$ identifiable? No
- Consider the following two populations of units:
  Population 1:

$$Pr(Y_{i1} = 6, Y_{i0} = 4) = 1/3, \quad Pr(Y_{i1} = 8, Y_{i0} = 6) = 1/3,$$

$$Pr(Y_{i1} = 10, Y_{i0} = 8) = 1/3$$

  Population 2:

$$Pr(Y_{i1} = 10, Y_{i0} = 4) = 1/3, \quad Pr(Y_{i1} = 8, Y_{i0} = 8) = 1/3,$$

$$Pr(Y_{i1} = 6, Y_{i0} = 6) = 1/3$$

Median treatment effect

- The distribution of treatment effects is:
  Population 1: $(2, 2, 2)$ with probability (1/3,1/3,1/3), hence the effect of the treatment is always 2!
  Population 2: $(6, 0, 0)$ with probability (1/3,1/3,1/3), hence the median treatment effect is 0
- The marginal distributions of Y_{i1} and Y_{i0} are the same in both populations
- **However** the treatment effect is determined by the joint distribution of (Y_{i1}, Y_{i0}), and the joint is different between the two populations
- Imagine the ideal experiment, can we ever observe the joint distribution of potential outcome? No, we can only extract certain moments and parameters, depending on the assumptions that the researcher makes on the DGP.

Median treatment effect: Another example

- Consider the following two populations:
  Population 1:

$$Pr(Y_{i1} = 1, Y_{i0} = 0) = 1/3, Pr(Y_{i1} = 3, Y_{i0} = 1) = 1/3,$$

$$Pr(Y_{i1} = 4, Y_{i0} = 3) = 1/3$$

  Population 2:

$$Pr(Y_{i1} = 4, Y_{i0} = 0) = 1/3, Pr(Y_{i1} = 3, Y_{i0} = 1) = 1/3,$$

$$Pr(Y_{i1} = 1, Y_{i0} = 3) = 1/3$$

- In population 1 the treatment effect is, $(1, 2, 1)$ and in population 2 the treatment effect is, $(4, 2, -2)$

Median treatment effect: Continuous variable example

- Let the joint distribution of the potential outcome be,

$$(Y_1, Y_0) \sim N((1, 0),\ \Sigma),$$

$$\Sigma = \begin{pmatrix} \mathbb{V}(Y_1) & Cov(Y_1, Y_0) \\ Cov(Y_1, Y_0) & \mathbb{V}(Y_0) \end{pmatrix}$$

- A binary treatment T is assigned at random.
- Can we identify the ATE? Can we identify the median treatment effect? can we identify percentiles of the treatment effect?

Median treatment effect: Continuous variable example

- Can we distinguish between this two distributions of the potential outcomes?
- Distribution 1,

$$\Sigma_1 = \begin{pmatrix} \mathbb{V}(Y_1) & Cov(Y_1, Y_0) \\ Cov(Y_1, Y_0) & \mathbb{V}(Y_0) \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

- Distribution 2,

$$\Sigma_1 = \begin{pmatrix} \mathbb{V}(Y_1) & Cov(Y_1, Y_0) \\ Cov(Y_1, Y_0) & \mathbb{V}(Y_0) \end{pmatrix} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$$

Median treatment effect: Continuous variable example

- Distribution 1,

$$\tau_1 = Y_1 - Y_0 \sim N(1, \mathbb{V}(Y_0) + \mathbb{V}(Y_1)) = N(1, 2)$$

- Distribution 2,

$$\tau_2 = Y_1 - Y_0 \sim N(1, \mathbb{V}(Y_0) + \mathbb{V}(Y_1) - 2Cov(Y_1, Y_0) = N(1, 1)$$

- The ATE is identified, and also the median treatment effect, as both τ_1 and τ_2 are symmetric distributions centred at 1 (the ATE and the median are equal).
  However all the other moments are not identified

SUTVA

Definition

No interference between units: the observation on one unit should be unaffected by the particular assignment of treatment to the other units.

- *No-interference* is the assumption that the allocation of treatment to unit i has no effect on the outcome of unit j for all i and j.
- How can we formally write the *no-interference* assumption?
- Let $\mathbf{T} = (T_1, \ldots, T_N)$ be the treatment assignment of all the units.
- The potential outcome of individual i is a function of the treatment assignment $\mathbf{T}$, i.e., it can depend on the assignment to treatment of units other than i,

$$Y_i(\mathbf{T}) = Y_i(T_1, T_2, \ldots, T_N)$$

- *No-interference* constrain the potential outcome of individual i to depend only on his assignment to treatment,

$$Y_i(\mathbf{T}) = Y_i(T_1, T_2, \ldots, T_N) = Y_i(T_i) \in \{Y_i(0), Y_i(1)\}$$

SUTVA

- SUTVA is a slightly stronger assumption than *no-interference*, hence SUTVA implies *no-interference*, and the opposite does not hold
- In this course we refer to SUTVA and *no-interference* as equivalent terms
- See Rosenbaum (2007) suggests a methods to conduct inference and identify causal relationships when SUTVA is violated.
- It is an excellent source for farther reading on SUTVA violations in experiments.
- The paper uses non-parametric exact inference methods.

SUTVA

- Consider a uniform randomized experiment with two strata, four units in the first strata and two units in the second strata, for 6 units in total. Half the units in each stratum receive treatment.
- There are 12 possible treatment assignments contained in the set Ω.

$$\Omega = \left\{ \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 0 \end{bmatrix} \right.$$
$$\left. \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 1 \end{bmatrix} \right\}.$$

Causal Effects without assuming SUTVA

- Without SUTVA, a causal effect is defined for every possible combination of the treatment assignment.
- The potential outcome for unit i might be $Y_{i100000000000}$ or $Y_{i010000000000}$, etc.
- How many potential outcomes will each unit have in a sample with N observation? 2^N
- Potential outcomes are still well defined when SUTVA is not satisfied!

SUTVA: Rubin (1986)

Statistics and Causal Inference: Comment: Which If's Have Causal Answers

- In a comment to Holland (1986) Rubin provides a formal definition of SUTVA.
- There are N units indexed by $u = 1, \ldots, N$, T treatments indexed by $t = 1, \ldots, T$, and an outcome variable Y_{tu}
- Rubin's definition: *"SUTVA is simply the a priori assumption that the value of Y for unit u when exposed to treatment t will be the same no matter what mechanism is used to assign treatment t to unit u and no matter what treatments the other units receive"* $\forall_t, \forall_u$
- Examples when SUTVA is violated:
  1. There exist unrepresented versions of treatments: *Y depends on which version of treatment t was received*
  2. interference between units: *the outcome, Y, of unit u depends on whether unit u' received treatment t or t'*

SUTVA: Rubin (1986)

Statistics and Causal Inference: Comment: Which If's Have Causal Answers

- Does the following statement has a causal meaning?
  If the females at firm f had been male, their starting salaries would have averaged 20% *higher*
  No, the statement is causal meaningless
- Rubin's answer:
  "the statement, by itself, is too vague to have a clear formulation satisfying SUTVA and thus is too vague to admit a clear causal answer. What are the units, treatments, and outcomes such that SUTVA is satisfied? I am not at all sure how to define anything except Y, which clearly involves starting salary"
- See Rubin (1986) for a variety of ways to make the statement have a causal meaning

SUTVA: Example I

- Assume the following DGP (data generating process):

$$Y_i = \alpha + \tau T_i + X_i \beta + \epsilon_i$$

- Is SUTVA satisfied in this model? Yes
- If $Cov(X_i, \epsilon_i) \neq 0$, X_i is endogenous. Is SUTVA satisfied? Yes

SUTVA: Example II

- Consider the following model of the treatment effect (multiplicative treatment effect)
$$Y_{i1} = \tau Y_{i0}$$
- What is the *ATE* effect?
Answer: $\mathbb{E}(Y_{i1} - Y_{i0}) = \mathbb{E}(\tau Y_{i0} - Y_{i0}) = \mathbb{E}(Y_{i0})(\tau - 1)$
- How can we estimate τ?
- One solution is to employ the following transformation on the data, *log*:
$$log(Y_{i1}) = \tau + log(Y_{i0})$$

- Now τ is the *ATE* of the treatment after the transformation, and can be estimated by the difference in means

SUTVA: Example II

- Prior to the *log* transformation, what is the variance of the potential outcomes with the treatment? Is it equal to the variance under control?

$$\mathbb{V}(Y_{i1}) = \tau^2 \mathbb{V}(Y_{i0})$$

- After the *log* transformation, the variance in both groups is the same,

$$\mathbb{V}(Y_{i1}) = \mathbb{V}(Y_{i0} + \tau) = \mathbb{V}(Y_{i0})$$

SUTVA: Example III

- The outcome of individual i is effected positively by the outcomes of other individuals.
- For example, one argument is that being next to better students will increase the outcomes of lower students. There is a positive externalities between students achievements.
- Consider the following model,

$$Y_i = \epsilon_i + \beta T_i + \sum_{j \neq i} \lambda \cdot Y_j$$

  where ϵ_i is a random variable.
- In this model the outcome of individual i is effected by the outcomes of individual j, and hence is also effected by individual j's treatment assignment.

$$Y_i(\mathbf{T}) \neq Y_i(T_i)$$

- No-interference is not satisfied, and hence also SUTVA is not satisfied.

Estimating ATE: Difference in means

The difference in means is an unbiased estimator of the ATE, when $(Y_{i1}, Y_{i0} \perp T_i)$,

$$\mathbb{E}\left(\frac{1}{m}\sum_{i=1}^{N} T_i Y_i - \frac{1}{N-m}\sum_{i=1}^{N}(1-T_i)Y_i\right) =$$

$$\frac{1}{m}\sum_{i=1}^{N}\mathbb{E}(Y_i T_i) - \sum_{i=1}^{N}\frac{1}{N-m}\mathbb{E}((1-T_i)Y_i) =$$

$$\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}(Y_{i1}|T_i=1) - \sum_{i=1}^{N-m}\frac{1}{N-m}\mathbb{E}(Y_{i0}|T_i=0) = ATE$$

$$\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}(Y_{i1}) - \sum_{i=1}^{N-m}\frac{1}{N-m}\mathbb{E}(Y_{i0}) = \mathbb{E}(Y_{i1}) - \mathbb{E}(Y_{i0})$$

$$\mathbb{E}(Y_{i1} - Y_{i0}) = ATE$$

Decomposing the difference in means estimator

- The difference in means can be decomposed to two components:
  - Causal effect - the average effect of assigning a unit to the treatment.
  - Selection - The effect of the dependence between the potential outcomes and the treatment assignment,

$$\Pr(Y(0)|T=1) \neq \Pr(Y(0)|T=0)$$

.

- The difference in means is,

$$\begin{aligned}
\mathbb{E}(Y|T=1) - \mathbb{E}(Y|T=0) &= \mathbb{E}(Y(1)|T=1) - \mathbb{E}(Y(0)|T=0) \\
&= [\mathbb{E}(Y(1)|T=1) - \mathbb{E}(Y(0)|T=1)] \\
&\quad + [\mathbb{E}(Y(0)|T=1) - \mathbb{E}(Y(0)|T=0)] \\
&= PATT + \underbrace{\mathbb{E}(Y(0)|T=1) - \mathbb{E}(Y(0)|T=0)}_{\text{Selection}}
\end{aligned}$$

- In an RCT $(Y(1), Y(0) \perp T) \Rightarrow \mathbb{E}(Y(0)|T=1) = \mathbb{E}(Y(0)|T=0)$.

Estimating ATE: Difference in means

- Are the following claims True or False?
  - When the potential outcome with and without the treatment are correlated $Cov(Y(1), Y(0)) \neq 0$ the difference in means will be a biased estimator. *False*
  - When the potential outcome are correlated with the treatment assignment $Cov(Y(1), T) \neq 0$ the difference in means will be a biased estimator. *True*
- The ATE is the first moment of the distribution of treatment effects. What is the second moment of the treatment effect distribution? $\mathbb{E}\left[(Y(1) - Y(0))^2\right]$.
- Can we identify the second moment of the treatment effect distribution? *No,* $\mathbb{E}\left[(Y(1) - Y(0))^2\right] = \mathbb{V}(Y(1)) + \mathbb{V}(Y(0)) - 2Cov(Y(1), Y(0))$, *and* $Cov(Y(1), Y(0))$ *is part of the joint distribution of* $Y(1)$ *and* $Y(0)$ *that is never observed.*
- Can we estimate $Cov(Y(1), Y(0))$? *We cannot estimate a parameter that is not identified!*

Random coefficient model

- The most basic model of treatment effects is a constant additive treatment effect,

$$Y_i(1) = Y_i(0) + \tau$$

- The random coefficient model is allows for heterogeneous treatment effects.
- Consider the following model,

$$Y_i = \beta_i T_i + \epsilon_i$$

where $\beta_i \sim N(10, 2)$ and $\epsilon_i \sim N(0, 1)$

- In this model are the potential outcomes random variables? *Yes.*
- Is SUTVA satisfies? *Yes, the potential outcomes of individual i are a function of T_i, and not $T_j \Rightarrow Y_i(\mathbf{T} = (T_1, \ldots, T_N)) = Y_i(T_i)$.*

Random coefficient model

- What is the sample ATE (SATE) and what is the population ATE (PATE)?

$$\text{SATE} = \frac{1}{N}\sum_{i=1}^{N}(Y_i(1) - Y_i(0)) = \frac{1}{N}\sum_{i=1}^{N}(Y_i(1) - Y_i(0))$$
$$= \frac{1}{N}\sum_{i=1}^{N}(\epsilon_i + \beta_i - \epsilon_i) = \frac{1}{N}\sum_{i=1}^{N}\beta_i$$
$$\text{PATE} = \mathbb{E}[\beta_i] = 10$$

- Is the SATE an unbiased estimator for the PATE?

$$\mathbb{E}[\text{SATE}] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}(Y_i(1) - Y_i(0))\right] = \frac{1}{N}\sum_{i=1}^{N}(\mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0))]$$
$$= \frac{1}{N}\sum_{i=1}^{N}(\mathbb{E}[\epsilon_i + \beta_i] - \mathbb{E}[\epsilon_i)] = 10 = \mathit{PATE}$$

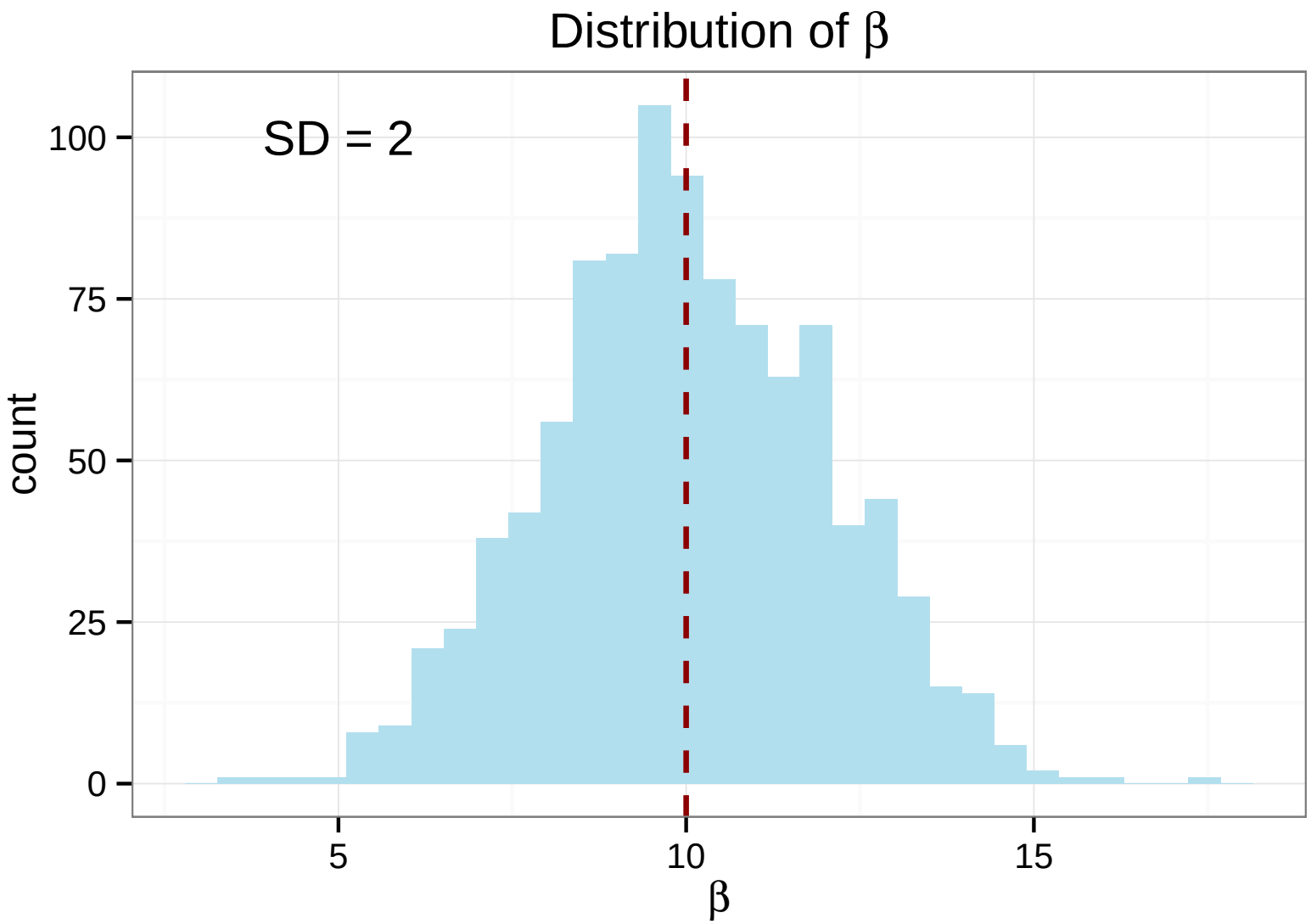
Random coefficient model: The efficiency gain of adjusting for covariates

- Consider the following model,

```
n=1000
beta = rnorm(n,mean=10,sd=2)
x = rexp(n,rate=0.1)
treat = rbinom(n,size=1,prob=0.5)
y = 5+beta*treat + x + rnorm(n,0,sd=1)
```

- We are interested in estimating PATE, $\mathbb{E}(\beta_i)$.
- What estimator should we use?
- What regressions should we perform?

Distribution of random coefficient



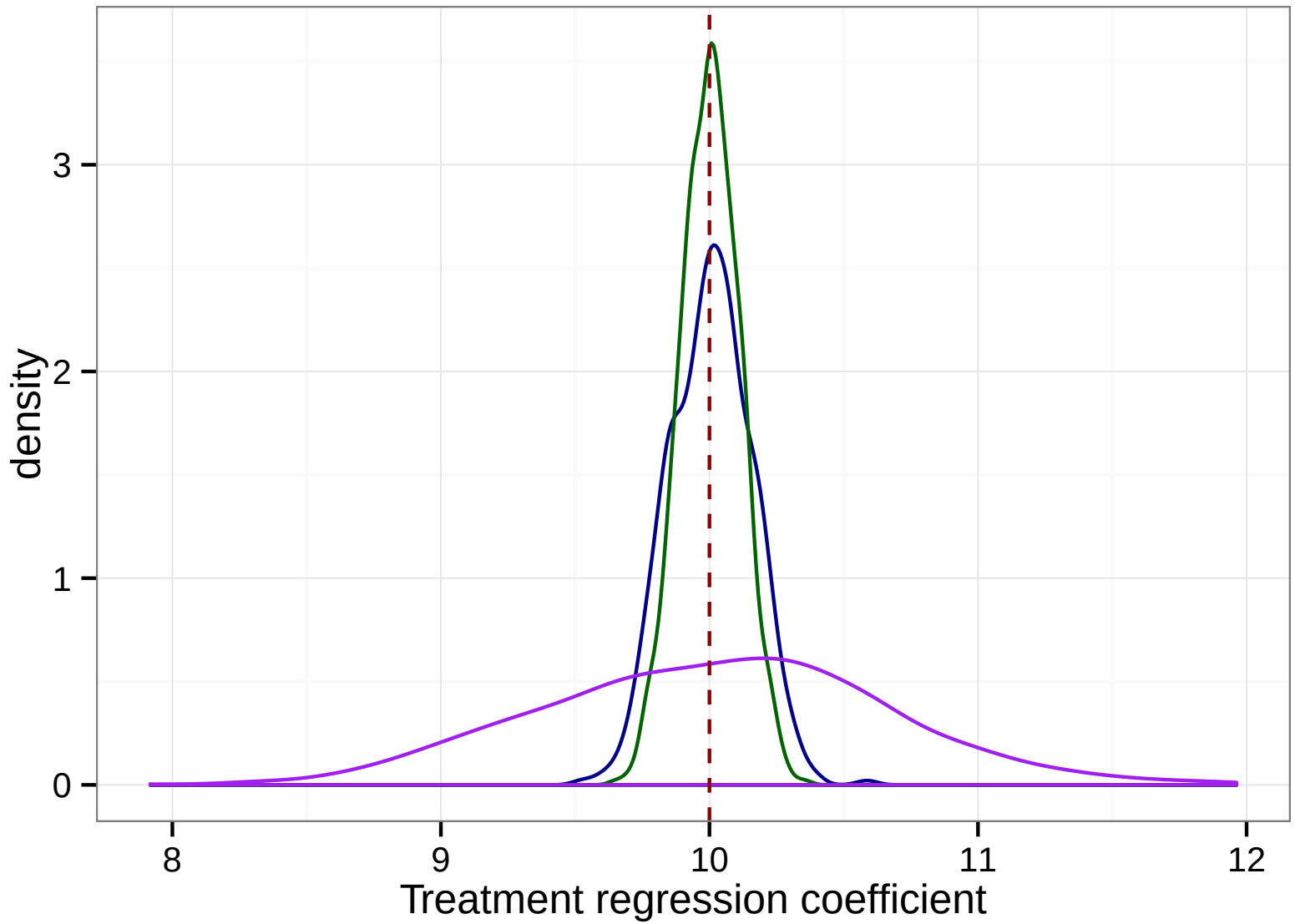
Random coefficient model: Regression results

	Model 1	Model 2	Model 3
(Intercept)	14.88***	5.03***	5.01***
	(0.46)	(0.09)	(0.11)
treat	10.60***	10.04***	10.08***
	(0.64)	(0.11)	(0.15)
x		1.00***	1.00***
		(0.01)	(0.01)
treat:x			−0.00
			(0.01)
R^2	0.21	0.98	0.98
Adj. R^2	0.21	0.98	0.98
Num. obs.	1000	1000	1000
RMSE	10.18	1.70	1.70

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

- In order to test the power of each of the models we samples 1,000 samples and estimated each one of the models in each sample.

Random coefficient model: Power simulation



Conditional independence assumption (CIA)

- The CIA implies that:

$$\mathbb{E}\left(Y_{i1}|X_i, T_i = 1\right) = \mathbb{E}\left(Y_{i1}|X_i, T_i = 0\right) = \mathbb{E}\left(Y_{i1}|X_i\right)$$

and

$$\mathbb{E}\left(Y_{i0}|X_i, T_i = 1\right) = \mathbb{E}\left(Y_{i0}|X_i, T_i = 0\right) = \mathbb{E}\left(Y_{i0}|X_i\right)$$

- Assuming CIA holds,

$$ATE = \mathbb{E}_{X_i}\left(\mathbb{E}_{Y_{i1}|X_i}\left(Y_{i1}|X_i, T_i = 1\right)\right) - \mathbb{E}_{X_i}\left(\mathbb{E}_{Y_{i0}|X_i}\left(Y_{i0}|X_i, T_i = 0\right)\right)$$

Conditional assumption (CIA)

- Assuming the following model (linear regression),

$$y_i = \alpha + \tau_1 T_i + X_i\beta + \epsilon$$

- Then,

$$\mathbb{E}(Y_i|T_i = 1, X_i) = \alpha + \tau_1 + X_i\beta, \;\; \mathbb{E}(Y_i|T_i = 0, X_i) = \alpha + X_i\beta$$

- In a regression model the standard assumption is that X_i is fixed (not a random variable), and therefore,

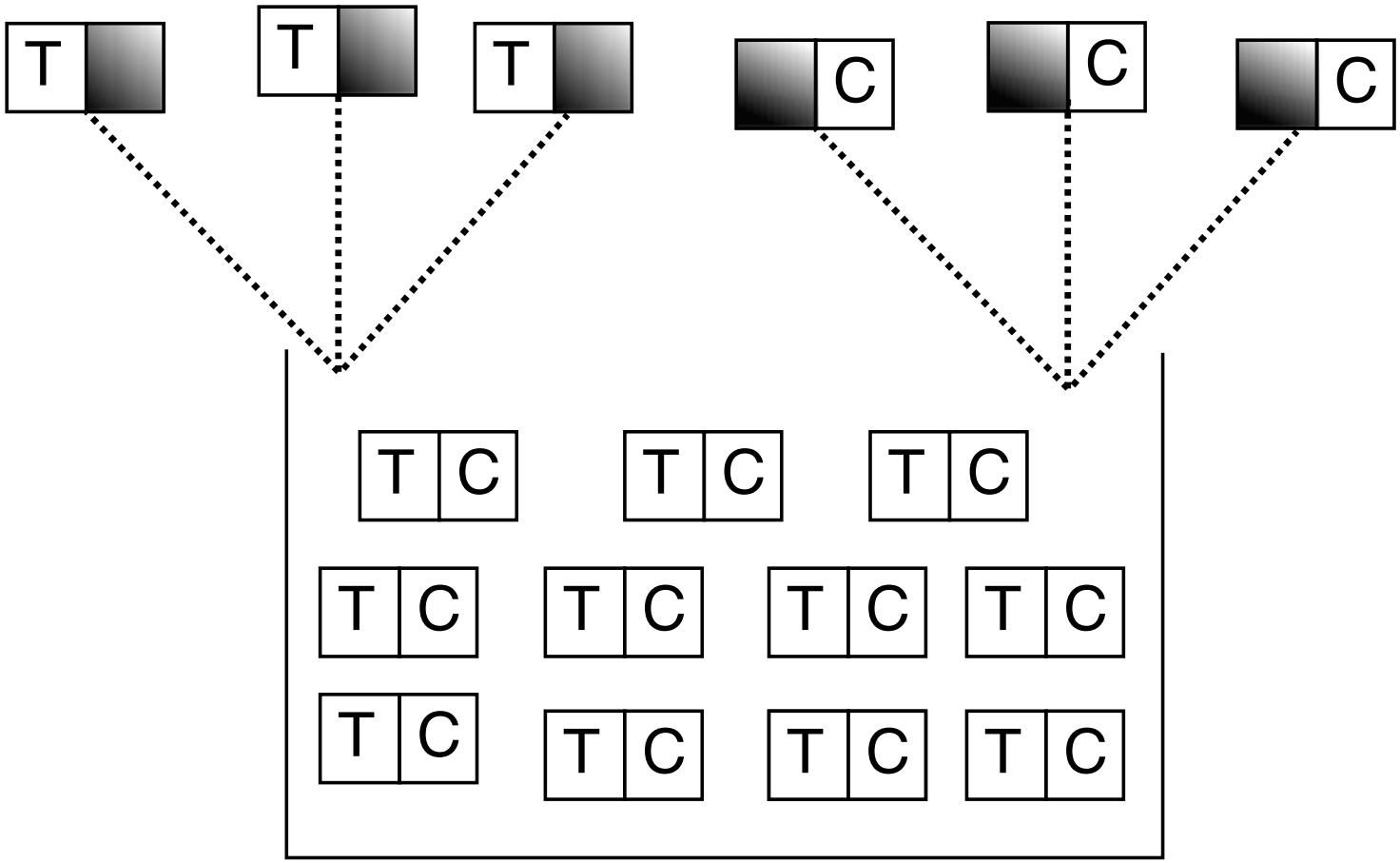
$$\mathbb{E}_{X_i}\left(\mathbb{E}_{Y_{i1}|X_i}\left(Y_{i1}|X_i, T_i = 1\right)\right) = \mathbb{E}_{Y_{i1}|X_i}\left(Y_{i1}|X_i, T_i = 1\right)$$

- Therefore the parameter τ_1 can be estimated by a regression adjustment, $\hat{\beta}^T_{OLS}$
- There are also many other ways of estimating τ_1, such as matching

Treatment assignment mechanisms

- There are many possible random treatment assignment mechanisms. The must common is selecting m observations to be assigned treatment out of N possible units
- In this approach, m, is fixed, it is not a random variable. The source of randomization is the random assignment of treatment

Treatment assignment mechanisms



Treatment assignment mechanisms

- There are N units, and m units are assigned a binary treatment at random
- Let Z_i be an indicator variable whether unit i was assigned treatment or control
- Is Z_i and Z_j independent? *No*
- What is $Cov(Z_i, Z_j) =$? Is it positive or negative? $Cov(Z_i, Z_j) < 0$, If unit i is assigned treatment the probability of unit j to receive treatment decreases. There is a negative relationship

Treatment assignment mechanisms

- What is, $Pr(Z_i = 1|m)$? $Pr(Z_i = 1|m) = \frac{m}{N}$
- Is Z_i and Z_j independent? What is $cov(Z_i, Z_j)$?
- When there are m units to be assigned treatment among N remaining units, the probability of $Z_i = 1$ conditional on Z_j is?
  $Pr(Z_i = 1|z_j = 0) = \frac{m}{N-1}$, $Pr(Z_i = 1|z_j = 1) = \frac{m-1}{N-1}$
- When $N \to \infty$: $Pr(Z_i = 1|z_j = 1) = Pr(Z_i = 1|z_j = 0) = Pr(Z_i = 1)$
- When $N \to \infty$, Z_i and Z_j are independent and $cov(Z_i, Z_j) = 0$

Calculating $Cov(Z_i, Z_j)$ Analytically

As Z_i is an indicator variable it follows that,

$$\mathbb{E}(Z_i) = Pr(Z_i = 1) = \frac{m}{N},\ \forall i, j$$

$$\mathbb{E}(Z_i \cdot Z_j) = 0 \times 0 \times Pr(Z_i = 0, Z_j = 0) + 1 \times 0 \times Pr(Z_i = 1, Z_j = 0) +$$
$$0 \times 1 \times Pr(Z_i = 0, Z_j = 1) + 1 \times 1 \times Pr(Z_i = 1, Z_j = 1)$$

$$= Pr(Z_i = 1, Z_j = 1) = \frac{m}{N} \cdot \frac{m-1}{N-1}$$

Hence,

$$Cov(Z_i, Z_j) = \mathbb{E}(Z_i \cdot Z_j) - \mathbb{E}(Z_i) \cdot \mathbb{E}(Z_j)$$
$$= \frac{m}{N}\left(\frac{m-1}{N-1} - \frac{m}{N}\right) < 0$$

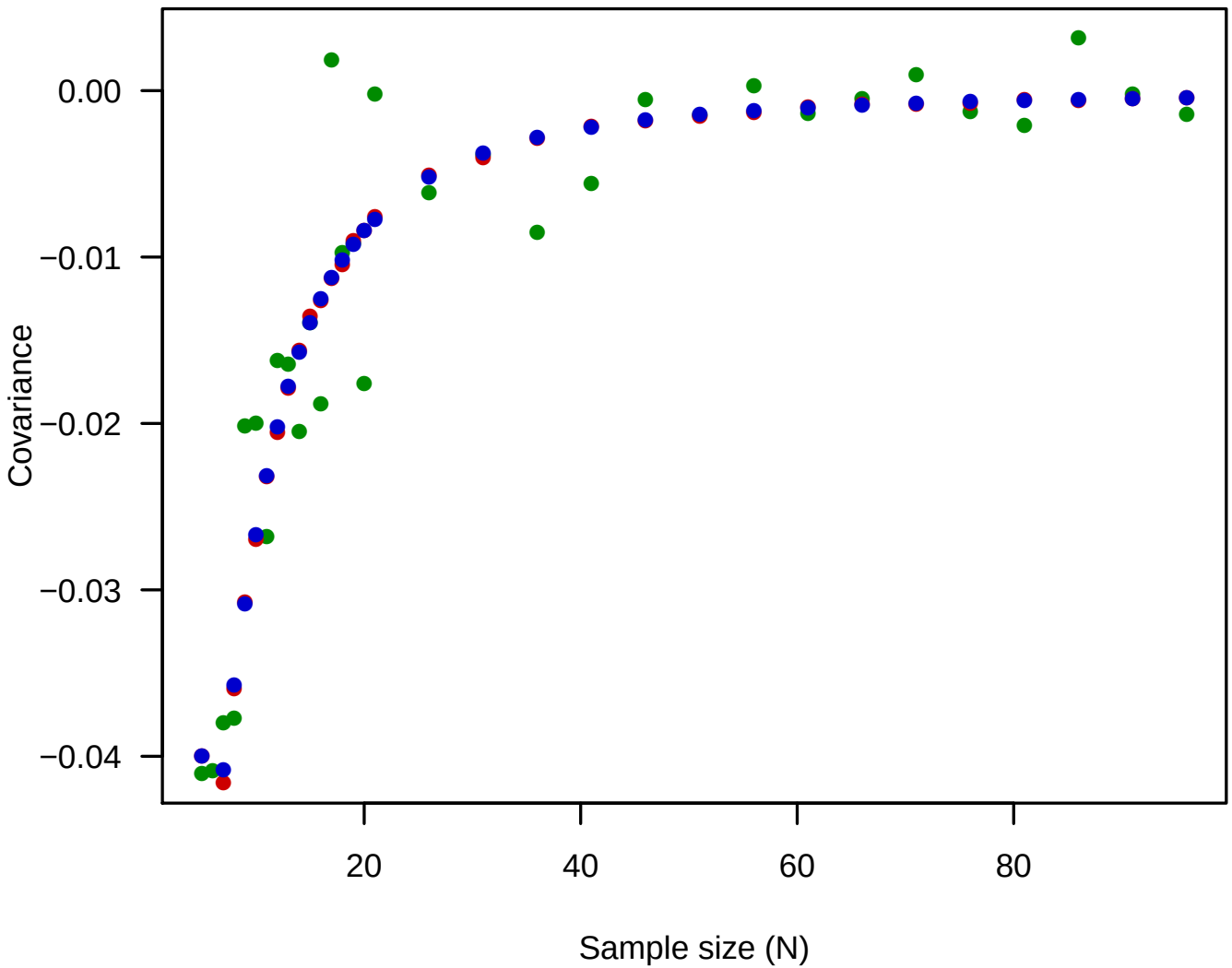
Monte Carlo simulations

- An alternative approach for estimating $Cov(Z_i, Z_j)$ is by a Monte-Carlo approximation
- The data generating process is known, a treatment was assigned at random, m units where chosen out of N. We can construct a simulation which performs exactly this process a multiple number of time and using the repetitions approximate the random component of the assignment mechanism.

Monte Carlo simulations: code

```
m=4
R=10000 #or 500000
n.vec = c(c(5:20),seq(21,100,by=5)) # sample sizes, N
cov.real1 <- cov.approx1 <- rep(999,length(n.vec))
for (i in c(1:length(n.vec))){
  N = n.vec[i]
 ## analytical:
 cov.real1[i] <- (m/N)*((m-1)/(N-1)-(m/N))
 ### Simulation:
  z1<-z2<-rep(999,R)
  for (j in c(1:R)){
    id.treat = sample(c(1:N),m,replace=FALSE)
    treat0 = rep(0,N)
    treat0[id.treat]=1
    z1[j] = treat0[1]
    z2[j] = treat0[2]
  }
  cov.approx1[i] <- cov(z1,z2)
}
```

Monte Carlo simulations: Results



R exercise

Write a code that generate the data $\{Y_t\}_{t=1}^{T}$:

$$Y_t = \theta Y_{t-1} + \epsilon_t, \quad \epsilon_t \sim N(0,1)$$

where $T = 100$.
Hint: Use a "while" loop.

R exercise: solution

```
# simulating data
num.T = 100
theta = 0.25
y0 = 0

y = rep(999,num.T)
y[1] = y0
t=1
while (t<num.T){
  y[t+1] =theta*y[t]+rnorm(1,mean=0,sd=1)
  t=t+1
}
```