Section 2 : Introduction to potential outcome and causal relationships
(and Monte – Carlo simulations)

Yotam Shem-Tov

Fall 2014
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,

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

<table>
<thead>
<tr>
<th>Group</th>
<th>$Y_{i1}$</th>
<th>$Y_{i0}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T = 1$</td>
<td>Observable: $Y_{i1}\mid T = 1$</td>
<td>Counterfactual: $Y_{i0}\mid T = 1$</td>
</tr>
<tr>
<td>$T = 0$</td>
<td>Counterfactual: $Y_{i1}\mid T = 0$</td>
<td>Observable: $Y_{i0}\mid T = 0$</td>
</tr>
</tbody>
</table>
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,

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

$$ATT = \mathbb{E}(Y_{i1} - Y_{i0} | T_i = 1)$$

$$ATC = \mathbb{E}(Y_{i1} - Y_{i0} | T_i = 0)$$

We can also be interested in the treatment effect conditional on a certain value of $Y_{i0}$, for example:

$$ATT' = \mathbb{E}(Y_{i1} - Y_{i0} | Y_{i0} \leq K)$$
**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 $P_\theta$ be a family of distributions indexed by $\theta$. A function of $\theta$ is identifiable if $f(\theta_1) \neq f(\theta_2)$ implies $P_{\theta_1} \neq P_{\theta_2}$ for all $\theta_1, \theta_2$.

**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,

$$E_p [g(x)] = (1 - p)g(0) + pg(1)$$

This is a linear function in $p$, however $\sqrt{p}$ is not a linear function of $p$. So, $E_p [g(x)] \neq \sqrt{p}$. 
Is the median treatment effect, \( \text{median}(Y_{i1} - Y_{i0}) \) identifiable? \textbf{No}

Consider the following two populations of units:

Population 1:

\[
\begin{align*}
Pr(Y_{i1} = 6, Y_{i0} = 4) &= 1/3, \\
Pr(Y_{i1} = 8, Y_{i0} = 6) &= 1/3, \\
Pr(Y_{i1} = 10, Y_{i0} = 8) &= 1/3
\end{align*}
\]

Population 2:

\[
\begin{align*}
Pr(Y_{i1} = 10, Y_{i0} = 4) &= 1/3, \\
Pr(Y_{i1} = 8, Y_{i0} = 8) &= 1/3, \\
Pr(Y_{i1} = 6, Y_{i0} = 6) &= 1/3
\end{align*}
\]
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? \textit{No}
Consider the following two populations:

Population 1:

\[ Pr(Y_{i1} = 1, Y_{i0} = 0) = \frac{1}{3}, Pr(Y_{i1} = 3, Y_{i0} = 1) = \frac{1}{3}, \]
\[ Pr(Y_{i1} = 4, Y_{i0} = 3) = \frac{1}{3} \]

Population 2:

\[ Pr(Y_{i1} = 4, Y_{i0} = 0) = \frac{1}{3}, Pr(Y_{i1} = 3, Y_{i0} = 1) = \frac{1}{3}, \]
\[ Pr(Y_{i1} = 1, Y_{i0} = 3) = \frac{1}{3} \]

In population 1 the treatment effect is, \((1, 2, 1)\) and in population 2 the treatment effect is, \((4, 2, -2)\)
Let the joint distribution of the potential outcome be,

\[(Y_1, Y_0) \sim N((1, 0), \Sigma),\]

\[
\Sigma = \begin{pmatrix}
\text{V}(Y_1) & \text{Cov}(Y_1, Y_0) \\
\text{Cov}(Y_1, Y_0) & \text{V}(Y_0)
\end{pmatrix}
\]

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?
Can we distinguish between this two distributions of the potential outcomes?

Distribution 1,

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

Distribution 2,

\[
\Sigma_1 = \begin{pmatrix}
V(Y_1) & Cov(Y_1, Y_0) \\
Cov(Y_1, Y_0) & V(Y_0)
\end{pmatrix} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}
\]
Median treatment effect: Continuous variable example

Distribution 1,

\[ \tau_1 = Y_1 - Y_0 \sim N(1, V(Y_0) + V(Y_1)) = N(1, 2) \]

Distribution 2,

\[ \tau_2 = Y_1 - Y_0 \sim N(1, V(Y_0) + V(Y_1) - 2Cov(Y_1, Y_0) = N(1, 1) \]

The ATE is identified, and also the median treatment effect, as both \( \tau_1 \) and \( \tau_2 \) are symmetric distributions centred at 1 (the ATE and the median are equal).

However all the other moments are not identified.
The difference in means is an unbiased estimator of the ATE, when 
\((Y_{i1}, Y_{i0} \perp T_i)\)

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

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

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

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

\[
\mathbb{E} (Y_{i1} - Y_{i0}) = ATE
\]
**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 $i, j$.
- 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.
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 $\Omega$. 

$$\Omega = \left\{ \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right\}$$
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!
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 $t'$
Does the following statement have 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
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 \( \text{Cov}(X_i, \epsilon_i) \neq 0 \), \( X_i \) is endogenous. Is SUTVA satisfied? Yes
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 \( \tau \)?

One solution is to employ the following transformation on the data, \( \log \):

\[ \log(Y_{i1}) = \tau + \log(Y_{i0}) \]

Now \( \tau \) is the ATE of the treatment after the transformation, and can be estimated by the difference in means.
Prior to the log transformation, what is the variance of the potential outcomes with the treatment? Is it equal to the variance under control?

$$\text{Var}(Y_{i1}) = \tau^2 \text{Var}(Y_{i0})$$

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

$$\text{Var}(Y_{i1}) = \text{Var}(Y_{i0} + \tau) = \text{Var}(Y_{i0})$$
The CIA implies that:

\[ \mathbb{E}(Y_{i1}|X_i, T_i = 1) = \mathbb{E}(Y_{i1}|X_i, T_i = 0) = \mathbb{E}(Y_{i1}|X_i) \]

and

\[ \mathbb{E}(Y_{i0}|X_i, T_i = 1) = \mathbb{E}(Y_{i0}|X_i, T_i = 0) = \mathbb{E}(Y_{i0}|X_i) \]

Assuming CIA holds,

\[
ATE = \mathbb{E}_{X_i} (\mathbb{E}_{Y_{i1}|X_i} (Y_{i1}|X_i, T_i = 1)) - \mathbb{E}_{X_i} (\mathbb{E}_{Y_{i0}|X_i} (Y_{i0}|X_i, T_i = 0))
\]
Conditional assumption (CIA)

- Assuming the following model (linear regression),
  \[ y_i = \alpha + \tau_1 T_i + X_i \beta + \epsilon \]

- Then,
  \[ \mathbb{E}(Y_i|T_i = 1, X_i) = \alpha + \tau_1 + X_i \beta, \quad \mathbb{E}(Y_i|T_i = 0, X_i) = \alpha + X_i \beta \]

- In a regression model the standard assumption is that \( X_i \) is fixed (not a random variable), and therefore,
  \[ \mathbb{E}_{X_i} \left( \mathbb{E}_{Y_i|X_i} (Y_i|X_i, T_i = 1) \right) = \mathbb{E}_{Y_i|X_i} (Y_i|X_i, T_i = 1) \]

- Therefore the parameter \( \tau_1 \) can be estimated by a regression adjustment, \( \hat{\beta}_{OLS}^T \)

- There are also many other ways of estimating \( \tau_1 \), such as matching
There are many possible random treatment assignment mechanisms. The most 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
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 $\text{Cov}(Z_i, Z_j) =$? Is it positive or negative?

$\text{Cov}(Z_i, Z_j) < 0$, If unit $i$ is assigned treatment the probability of unit $j$ to receive treatment decreases. There is a negative relationship.
What is, \( Pr(Z_i = 1|m) \)? \( Pr(Z_i = 1|m) = \frac{m}{N} \)

Is \( Z_i \) and \( Z_j \) independent? What is \( \text{cov}(Z_i, Z_j) \)?

When there are \( m \) units to be assigned treatment among \( N \) remaining units, the probability of \( Z_i = 1 \) conditional on \( Z_j \) is:
\[
Pr(Z_i = 1|z_j = 0) = \frac{m}{N-1}, \quad Pr(Z_i = 1|z_j = 1) = \frac{m-1}{N-1}
\]

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 \( \text{cov}(Z_i, Z_j) = 0 \)
Calculating $\text{Cov}(Z_i, Z_j)$ 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,

$$\text{Cov}(Z_i, Z_j) = \mathbb{E}(Z_i \cdot Z_j) - \mathbb{E}(Z_i) \cdot \mathbb{E}(Z_j)$$

$$= \frac{m}{N} \left( \frac{m - 1}{N - 1} - \frac{m}{N} \right) < 0$$
An alternative approach for estimating $\text{Cov}(Z_i, Z_j)$ is by a Monte-Carlo approximation.

The data generating process is known, a treatment was assigned at random, $m$ units were 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

Sample size (N) vs. Covariance