CAUSAL INFERENCE - Example Sheet 1 Solutions

J. Hera Shi, Part III Michaelmas 2023

Q 1 Let $B_i = B_i(x_{[n]})$ be the group (stratum) of unit i.

$$\pi(a_{[n]}|x_{[n]}) = \begin{cases} \prod_{j=1}^{m} \binom{n_j}{n_{1j}}^{-1}, & \text{if } \sum_{i=1}^{n} a_i I(B_i = j) = n_{1j} \text{ and } \sum_{i=1}^{n}(1-a_i)I(B_i = j) = n_j - n_{1j} \text{ for all } j \in [n] \\ 0, & \text{otherwise.} \end{cases}$$

Q 2 It's easy to show that $F \circ F^{-1}(\alpha) \geq \alpha$ for all $\alpha \in [0,1]$. We have that if $U \sim \text{Unif}[0,1]$, then $F^{-1}(U) \sim F$. Therefore

$$\begin{aligned} \mathbb{P}(F(T) \leq \alpha) &= \mathbb{P}(F \circ F^{-1}(U) \leq \alpha) \\ &\leq \mathbb{P}(U \leq \alpha) \\ &= \alpha \end{aligned}$$

Q 3 We denote $n_1 = \sum_i^n A_i$ and $n_0 = \sum_{i=1}^n (1 - A_i)$. We drop conditioning on $\mathbf{W}$ in the following derivation for notational convenience.

Step 1. Rewrite the estimator as:

$$\begin{aligned} \hat{\beta}_n &= \frac{1}{n_1}\sum_{i=1}^{n} A_i Y_i - \frac{1}{n_0}\sum_{i=1}^{n}(1-A_i)Y_i \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{n}{n_1}A_i Y_i^1 - \frac{n}{n_0}(1-A_i)Y_i^0\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{n}{n_1}A_i Y_i^1 - \frac{n}{n_0}(\frac{n_1+n_0}{n} - A_i)Y_i^0\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{n}{n_1}A_i Y_i^1 - \frac{n}{n_0}(\frac{n_1+n_0}{n} - A_i)Y_i^0\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{n}{n_1}\left(\underbrace{A_i - \frac{n_1}{n}}_{Z_i} + \frac{n_1}{n}\right)Y_i^1 - \frac{n}{n_0}\left(\frac{n_0}{n} - \left(\underbrace{A_i - \frac{n_1}{n}}_{Z_i}\right)\right)Y_i^0\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}(Y_i^1 - Y_i^0) + \frac{1}{n}\sum_{i=1}^{n} Z_i\left(\frac{n}{n_1}Y_i^1 + \frac{n}{n_0}Y_i^0\right) \\ &= \beta + \frac{1}{n}\sum_{i=1}^{n} Z_i\left(\frac{n}{n_1}Y_i^1 + \frac{n}{n_0}Y_i^0\right) \end{aligned}$$

Step 2. Thus far, we have shown that the difference-in-means estimator is equivalent to:

$$\hat{\beta}_n = \beta_n + \frac{1}{n}\sum_{i=1}^{n} Z_i\left(\frac{n}{n_1}Y_i^1 + \frac{n}{n_0}Y_i^0\right), \tag{1}$$

in which the only random quantities are $\{Z_i\}_{i=1}^n$. Moreover, $\mathbb{E}[Z_i] = 0$, $\text{var}[Z_i] = \frac{n_1 n_0}{n}$, and $\text{cov}[Z_i Z_j] = \mathbb{E}[Z_i Z_j] = \mathbb{E}[A_i A_j] - \left(\frac{n_1}{n}\right)^2 = -\frac{n_1 n_0}{n(n-1)}$ for $i, j \in [n], i \neq j$.

Step 3. Expand the variance form using the expectation, variance, and covariance we had in **Step 2**:

$$\begin{aligned}
\text{var}(\hat{\beta}_n) &= \text{var}\left(\beta_n + \frac{1}{n}\sum_{i=1}^{n} Z_i\left(\frac{n}{n_1}Y_i^1 + \frac{n}{n_0}Y_i^0\right)\right) \\
&= \text{var}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i\left(\frac{n}{n_1}Y_i^1 + \frac{n}{n_0}Y_i^0\right)\right) \\
&= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i\left(\frac{n}{n_1}Y_i^1 + \frac{n}{n_0}Y_i^0\right)\right)^2 \\
&= \sum_{i=1}^{n}\left[\frac{1}{nn_1} - \frac{n_1-1}{n(n-1)n_1}\right](Y_i^1)^2 + \sum_{i=1}^{n}\left[\frac{1}{nn_0} - \frac{n_0-1}{n(n-1)n_0}\right](Y_i^0)^2 + \sum_{i=1}^{n}\frac{2}{n(n-1)}Y_i^1Y_i^0 \\
&= \frac{1}{n_1}S(1)^2 + \frac{1}{n_0}S(0)^2 - \frac{1}{n}S(0,1)^2.
\end{aligned}$$

A bit of reflection Difference-in-means estimator is unbiased for the sample average treatment effect:

$$\begin{aligned}
\mathbb{E}[\hat{\beta}] &= \mathbb{E}\left[\beta_n + \frac{1}{n}\sum_{i=1}^{n} Z_i\left(\frac{n}{n_1}Y_i^1 + \frac{n}{n_0}Y_i^0\right)\right] \\
&= \beta_n,
\end{aligned}$$

where $\mathbb{E}[Z_i] = 0$ for all $i \in [n]$. The following expressions are unbiased estimators for S_0^2 and S_1^2, respectively.

$$\hat{S}_0^2 = \frac{1}{n_0 - 1}\sum_{i=1}^{n}(1 - A_i)(Y_i - \bar{Y}_0)^2 \qquad \hat{S}_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n} A_i(Y_i - \bar{Y}_1)^2$$

The last term in the estimation expression can be written as:

$$\begin{aligned}
S(0,1)^2 &= \frac{1}{n-1}\sum_{i=1}^{n}\{Y_i^1 - Y_i^0 - \beta_n\}^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\{Y_i^1 - Y_i^0 - \beta_n\}^2
\end{aligned}$$

Q 4 We use the notation $\pi := \mathbb{P}(A_i = 1|X_i)$. First, we consider the most general case and set the derivatives w.r.t. γ_3 and δ_3 to zero

$$\frac{\partial}{\partial\gamma_3}\mathbb{E}\left[(Y - \alpha_3 - \beta_3 A - \gamma_3^T X - A(\delta_3^T X))^2\right] = \mathbb{E}[XY] - \gamma_3\mathbb{E}[XX^T] - \pi\delta_3\mathbb{E}[XX^T] = 0, \tag{2}$$

$$\frac{\partial}{\partial\delta_3}\mathbb{E}\left[(Y - \alpha_3 - \beta_3 A - \gamma_3^T X - A(\delta_3^T X))^2\right] = \mathbb{E}[AXY] - \pi\gamma_3\mathbb{E}[XX^T] - \pi\delta_3\mathbb{E}[XX^T] = 0. \tag{3}$$

Subtracting (3) from (2), we get

$$\mathbb{E}[XY] - \mathbb{E}[AXY] - (1-\pi)\gamma_3\mathbb{E}[XX^T] = 0 \quad \Leftrightarrow \quad \gamma_3 = \frac{\mathbb{E}[XX^T]^{-1}\mathbb{E}[(1-A)XY]}{1-\pi}.$$

We cancel $1-\pi$ and apply the tower property as well as Theorem 2.12 to get an expression in terms of the counterfactual.

$$\begin{aligned}
\gamma_3 &= \frac{\mathbb{E}[XX^T]^{-1}\mathbb{E}[XY|A=0](1-\pi)}{1-\pi} = \mathbb{E}[XX^T]^{-1}\mathbb{E}[XY|A=0] \\
&= \mathbb{E}[XX^T]^{-1}\mathbb{E}\big[\mathbb{E}[XY|A=0,X]\big] = \mathbb{E}[XX^T]^{-1}\mathbb{E}\big[X\,\mathbb{E}[Y(0)|X]\big] = \mathbb{E}[XX^T]^{-1}\mathbb{E}[Y(0)\,X].
\end{aligned}$$

From (3), we get

$$\delta_3 + \gamma_3 = \frac{\mathbb{E}[XX^T]^{-1}\mathbb{E}[AXY]}{\pi} \quad \Leftrightarrow \quad \delta_3 = \frac{\mathbb{E}[XX^T]^{-1}\mathbb{E}[XY|A=1]\pi}{\pi} - \gamma_3.$$

By a similar reasoning as above, we arrive at

$$\delta_3 = \mathbb{E}[XX^T]^{-1}\big(\mathbb{E}[XY|A=1] - \mathbb{E}[XY|A=0]\big) = \mathbb{E}[XX^T]^{-1}\mathbb{E}[X(Y(1)-Y(0))].$$

For γ_2, we consider equation (2) and replace γ_3 by γ_2 and δ_3 by 0. Therefore, we get

$$\begin{aligned}\gamma_2 &= \mathbb{E}[XX^T]^{-1}\mathbb{E}[XY] = \mathbb{E}[XX^T]^{-1}\mathbb{E}\Big[X\big(\pi\mathbb{E}[Y(1)|X] + (1-\pi)\mathbb{E}[Y(0)|X]\big)\Big] \\ &= \mathbb{E}[XX^T]^{-1}\big(\pi\mathbb{E}[Y(1)\,X] + (1-\pi)\mathbb{E}[Y(0)\,X]\big)\end{aligned}$$

Q 5 We prove the most general case, i.e. $m = 3$, and introduce the notations $Z = (1, A, X, AX)^T$, $\theta = (\alpha_3, \beta_3, \gamma_3, \delta_3)^T$, $\Sigma = XX^T$ and $\varepsilon_3 = Y - \theta^T Z$. Recall $\mathbb{E}[X] = 0$, $A \perp\!\!\!\perp X$, $\pi = \mathbb{P}(A=1)$ and $A^2 = A$. First, we use these relationships to find an expression for $\mathbb{E}[ZZ^T]$:

$$\mathbb{E}[ZZ^T] = \ldots = \begin{pmatrix} 1 & \pi & 0 & 0 \\ \pi & \pi & 0 & 0 \\ 0 & 0 & \Sigma & \pi\Sigma \\ 0 & 0 & \pi\Sigma & \pi\Sigma \end{pmatrix}.$$

From Lemma 2.26, we know that

$$V_3 = \Big[\mathbb{E}[ZZ^T]^{-1}\mathbb{E}[ZZ^T\varepsilon_3^2]\mathbb{E}[ZZ^T]^{-1}\Big]_{22}.$$

Since $\mathbb{E}[ZZ^T]$ has a block structure, the formula for V_3 reduces to

$$V_3 = \left[\begin{pmatrix} 1 & \pi \\ \pi & \pi \end{pmatrix}^{-1} \mathbb{E}\begin{pmatrix} \varepsilon_3^2 & A\varepsilon_3^2 \\ A\varepsilon_3^2 & A\varepsilon_3^2 \end{pmatrix} \begin{pmatrix} 1 & \pi \\ \pi & \pi \end{pmatrix}^{-1}\right]_{22} = \ldots = \frac{\mathbb{E}[(A-\pi)^2\varepsilon_3^2]}{\pi^2(1-\pi)^2}.$$

The derivation of V_1 and V_2 works analogously by replacing Z with $(1, A)^T$ and $(1, A, X)^T$, respectively.
The formula we have just derived yields that $V_1 = V_2 = V_3$ if and only if $\mathbb{E}[(A-\pi)^2\varepsilon_1] = \mathbb{E}[(A-\pi)^2\varepsilon_2] = \mathbb{E}[(A-\pi)^2\varepsilon_3]$. From the proof of Theorem 2.30, we know that

$$\begin{aligned}\varepsilon_1 &= \varepsilon_3 + \gamma_3^T X + A(\delta_3^T X), \\ \varepsilon_2 &= \varepsilon_3 + (\gamma_3 - \gamma_2)^T X + A(\delta_3^T X),\end{aligned}$$

which implies that $\gamma_2 = \gamma_3 = \delta_3 = 0 \Rightarrow V_1 = V_2 = V_3$.
We construct a counter-example to disprove $V_2 < V_1$. From the equations above, we get $\varepsilon_1 = \varepsilon_2 + \gamma_2^T X$. Moreover, applying the formula for the variance yields

$$V_1 \le V_2 \;\Leftrightarrow\; \mathbb{E}[(A-\pi)^2\varepsilon_1^2] - \mathbb{E}[(A-\pi)^2\varepsilon_2] = \mathbb{E}[(A-\pi)^2(2\gamma_2^T X\varepsilon_2 + \gamma_2^T XX^T\gamma_2)] \le 0. \qquad (4)$$

We use the data-generating mechanism $Y = -\frac{3}{4}X + AX$ with $A \sim \text{Bernoulli}(1/4)$ and estimate γ_2 as

$$\gamma_2 = \frac{\mathbb{E}[XY]}{\mathbb{E}[X^2]} = \frac{\mathbb{E}[-\frac{3}{4}X^2 + AX^2]}{\mathbb{E}[X^2]} = -\frac{1}{2}.$$

Hence, we obtain $\varepsilon_2 = Y - \gamma_2 X = -\frac{1}{4}X + AX$ and insert it into (4)

$$\mathbb{E}\left[\Big(A - \frac{1}{4}\Big)^2\Big(\frac{1}{4}X^2 - AX^2 + \frac{1}{4}X^2\Big)\right] = \mathbb{E}[X^2]\mathbb{E}\left[\Big(A - \frac{1}{4}\Big)^2\Big(\frac{1}{2} - A\Big)\right] = -\frac{1}{32}\mathbb{E}[X^2] < 0.$$

Remark. The randomization assumption is not violated by the counterexample. By consistency, we have $Y(A) = Y = -\frac{3}{4}X + AX$ and, hence, $Y(1) = -\frac{3}{4}X + X$. Then, $Y(1) \perp\!\!\!\perp A$ follows from $A \perp\!\!\!\perp X$ and $Y(0) \perp\!\!\!\perp A$ works analogously.

Q 6 The conditional distribution of N_{11} given the column margins $N_{\cdot 0}$ and $N_{\cdot 1}$ is:

$$\begin{aligned} N_{01} &\sim \text{Bin}(N_{0\cdot}, \pi_0) \\ N_{11} &\sim \text{Bin}(N_{1\cdot}, \pi_1) \\ N_{\cdot 1} &= N_{01} + N_{11} \end{aligned}$$

Therefore, the conditional density under the null (i.e. $\pi_0 = \pi_1 = \pi$) can be written as:

$$\begin{aligned} \mathbb{P}(N_{11}|N_{\cdot 1}, N_{\cdot 0}) &= \frac{\binom{N_{1\cdot}}{N_{11}} \pi^{N_{11}}(1-\pi)^{N_{10}} \binom{N_{0\cdot}}{N_{01}} \pi^{N_{01}}(1-\pi)^{N_{00}}}{\binom{N_{\cdot\cdot}}{N_{\cdot 1}} \pi^{N_{\cdot 1}}(1-\pi)^{N_{\cdot 0}}} \\ &= \frac{\binom{N_{1\cdot}}{N_{11}}\binom{N_{0\cdot}}{N_{01}}}{\binom{N_{\cdot\cdot}}{N_{\cdot 1}}} = \frac{\binom{N_{1\cdot}}{N_{10}}\binom{N_{0\cdot}}{N_{00}}}{\binom{N_{\cdot\cdot}}{N_{\cdot 0}}} \end{aligned}$$

which takes the same form as Fisher's exact test in the randomization model shown in the lecture notes.

Q 7 (a) If $\mathbf{A}$ is randomized by sampling without replacement, the treatment assignments are exchangeable, that is

$$(A_1, \ldots, A_j, \ldots, A_n) \overset{\text{D}}{=} (A_{\sigma(1)}, \ldots, A_{\sigma(j)}, \ldots, A_{\sigma(n)})$$

for any permutation σ. Define $\psi_i := \psi(i/n)$ and $\psi_i^r := \psi(r_i/n)$ for $i \in [n]$. Conditional on $\mathbf{W}$, the ψ_i^r are fixed and there exists a permutation σ such that $\psi_i = \psi^r_{\sigma(i)}$ for all $i \in [n]$. Hence,

$$\begin{aligned} \mathbb{P}(T(\mathbf{A}, \mathbf{W}) \le t|\mathbf{W}) &= \mathbb{P}\Big(\sum_{i=1}^n A_i \psi_i^r \le t \Big| \mathbf{W}\Big) = \mathbb{P}\Big(\sum_{i=1}^n A_{\sigma(i)} \psi_i \le t \Big| \mathbf{W}\Big) \\ &= \mathbb{P}\Big(\sum_{i=1}^n A_i \psi_i \le t \Big| \mathbf{W}\Big) = \mathbb{P}\Big(\sum_{i=1}^n A_i \psi_i \le t\Big), \end{aligned}$$

where the penultimate equality follows from exchangeability and the last equality follows from $\mathbf{A} \perp\!\!\!\perp \mathbf{W}$. Hence, conditional on $\mathbf{W}$, the distribution of $T(\mathbf{A}, \mathbf{W})$ does not depend on the ranks $\mathbf{W}$.

(b) When $\psi(r) = r$, we are looking at a Wilcoxon rank sum statistic defined as:

$$T(\mathbf{A}, \mathbf{W}) = \sum_{i=1}^n A_i \frac{r_i}{n} = \frac{1}{n} \sum_{i=1}^n A_i r_i$$

Recall that $1 + 2 + \ldots + k = k(k+1)/2$. If the treated subjects (assume total n_1 units) have the smallest possible ranks, 1 to n_1, then

$$T = \frac{1}{n}\left(1 + 2 + \ldots + n_1\right) = \frac{n_1(n_1+1)}{2n}.$$

If the treated subjects have the largest possible ranks, $n - n_1 + 1$ to n, then

$$T = \frac{1}{n}\left(n - n_1 + 1 + n - n_1 + 2 + \cdots + n\right) = \frac{n_1(2n - n_1 + 1)}{2n}.$$

All the integers between $n_1(n_1+1)/2$ and $n_1(2n - n_1 + 1)$ are possible values of nT. The null distribution of T under the sharp null hypothesis is symmetric about $\frac{n_1(n+1)}{2n}$, and thus:

$$\bar{T} = \frac{n_1(n+1)}{2n} \tag{5}$$

Another way to think about this problem is by using the linearity of expectation, which tells us that

$$\bar{T} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[A_i r_i] = \mathbb{E}[A_i r_i] = \mathbb{E}[A_i]\mathbb{E}[r_i] = \frac{n_1(n+1)}{2n}$$

Now we consider the **variance**. This follows from the fact that the variance of the sample sum of a simple random sample of size n_1 from a list of n numbers is:

$$\mathrm{var}(nT) = \frac{(n-n_1)n_1\,\mathrm{var}(r_i)}{n-1}$$

where r_i is uniformly distributed, thus $\mathrm{var}(r_i) = \frac{(n-1)(n+1)}{12}$. In conclusion:

$$\mathrm{var}(T) = \frac{(n-n_1)n_1(n+1)}{12n^2}$$

By CLT, when n is "large enough", the distribution of test statistics T could be approximated by a normal distribution:

$$\mathcal{N}\left(\frac{n_1(n+1)}{2n}, \frac{(n-n_1)n_1(n+1)}{12n^2}\right) \tag{6}$$

Equivalently:

$$\frac{T_n - \frac{n_1(n+1)}{2n}}{\sqrt{\frac{(n-n_1)n_1(n+1)}{12n^2}}} \sim^d \mathcal{N}(0,1) \tag{7}$$

We can base a test of the sharp null hypothesis on T. To test against the alternative that treatment tends to increase responses, we would reject large values of T. To test against the alternative that treatment tends to decrease responses, we would reject small values of T. The critical value of the test is set using the probability distribution of T on the assumption that the sharp null hypothesis is true. For a level-alpha test against the alternative that treatment increases responses, we would find the smallest c such that, if the sharp null is true, $P(T \geq c) \leq \alpha$. We will reject the sharp null if the observed value of T is c or greater.

Q 8 (a) First, we compute E for the difference-in-means estimator:

$$\mathbb{E}\big[T(\mathbf{A},\mathbf{X},\mathbf{Y}(0))|\mathbf{X},\mathbf{Y}(0)\big] = \mathbb{E}\left[\frac{1}{n_1}\sum_{i=1}^{n}A_iY_i(0) - \frac{1}{n-n_1}\sum_{i=1}^{n}(1-A_i)Y_i(0)\Big|\mathbf{X},\mathbf{Y}(0)\right] = 0,$$

since $\mathbb{E}[A_i] = \frac{n_1}{n}$ for all $i \in [n]$. According to the definition, the Hodges-Lehmann estimator needs to fulfill

$$\frac{1}{n_1}\sum_{i}^{n}A_i(Y_i - \hat{\beta}_{\mathrm{HL}}A_i) - \frac{1}{n_0}\sum_{i=1}^{n}(1-A_i)(Y_i - \hat{\beta}_{\mathrm{HL}}A_i) = 0.$$

As $A_i(1-A_i) = 0$ and $\sum_{i=1}^{n}A_i^2 = n_1$, we arrive at

$$\hat{\beta}_{\mathrm{HL}} = \frac{1}{n_1}\sum_{i}^{n}A_iY_i - \frac{1}{n_0}\sum_{i=1}^{n}(1-A_i)Y_i.$$

(b) Let j_1 and j_2 be the indices of the two observations belonging to the pair $D_j, j \in [m]$. By definition of the treatment assignment mechanism, we have $D_j = (A_{j_1} - A_{j_2})(Y_{j_1} - Y_{j_2})$ for $j \in [m]$. Analogously to (a), we compute E for the sign statistic:

$$\begin{aligned}\mathbb{E}\big[T(\mathbf{A},\mathbf{X},\mathbf{Y}(0))\big|\mathbf{X},\mathbf{Y}(0)\big] &= \mathbb{E}\Big[\sum_{j=1}^{m}\mathrm{sign}\left((A_{j_1} - A_{j_2})(Y_{j_1}(0) - Y_{j_2}(0))\right)\Big|\mathbf{X},\mathbf{Y}(0)\Big] \\ &= \sum_{j=1}^{m}\mathbb{E}[\mathrm{sign}(A_{j_1} - A_{j_2})] = 0,\end{aligned}$$

as for every pair D_j, $A_{j_2} = 1 - A_{j_1}$ and, thus, $\mathbb{P}(A_{j_1} - A_{j_2} = 1) = \mathbb{P}(A_{j_1} - A_{j_2} = -1) = \frac{1}{2}$. It is easy to check that $T(\mathbf{A}, \mathbf{X}, \mathbf{Y} - \beta\mathbf{A}) = \sum_{j=1}^{m} \text{sign}(D_j - \beta)$. Therefore, the Hodges-Lehmann estimator is given by

$$\hat{\beta}_{\text{HL}} = \begin{cases} D_{(\frac{m+1}{2})}, & \text{if } m \text{ is odd,} \\ \frac{1}{2}D_{(\frac{m}{2})} + \frac{1}{2}D_{(\frac{m}{2}+1)}, & \text{if } m \text{ is even.} \end{cases}$$

(c) This directly follows from Theorem 2.19:

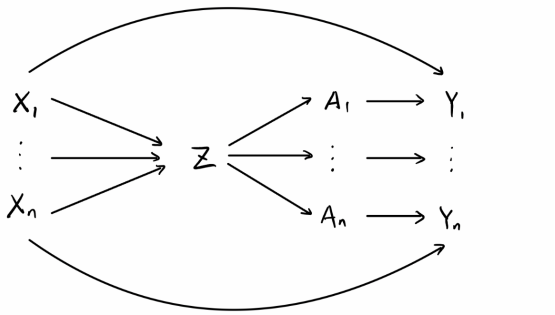
$$\mathbb{P}(\beta \in \mathcal{C}_\alpha) = \mathbb{P}(P(\beta) \geq \alpha) \geq 1 - \alpha.$$

If the randomization distribution $T(\mathbf{A}, \mathbf{X}, \mathbf{Y}(0))|\mathbf{W}$ does not depend on $\mathbf{W}$, then it has the same distribution for every choice of β in the null hypothesis. Hence, if we test a range of different β-values, we need to compute the quantiles of the distribution only once instead of for every β individually. This lowers the computational costs considerably.

Q 9 For a directed graph $\mathcal{G} = ([p], E)$ we prove: acyclicity $\Leftrightarrow$ there is a topological ordering of the nodes.
$\Leftarrow$ by contradition: Suppose there exists a topological ordering of $\mathcal{G}$, i.e. a permutation of $(k_1, \ldots, k_p)$ of $(1, \ldots, p)$ such that $(i, j) \in E$ implies $k_i < k_j$. Further, assume that the graph is not acyclic, that is there is a cycle $k_i \to k_j \to \ldots \to k_i$. This leads to the contradiction $k_i < k_i$.
$\Rightarrow$ by induction over the vertices:

- Base case $p = 1$: obviously true
- Induction hypothesis: For any DAG with l vertices, there is a topological ordering.
- Induction step $l \to l + 1$: Any DAG with $l + 1$ vertices has at least one vertex without incoming edges. Take one such vertex and call it k_1. Define the graph $\mathcal{G}' = (V', E')$ by deleting k_1 and the associated edges from $\mathcal{G}$. By the induction hypothesis, $\mathcal{G}'$ has a topological ordering$(k_2, \ldots, k_p)$. Hence, $(k_1, \ldots, k_p)$ is a topological ordering of the original graph $\mathcal{G}$.

We prove that for every $J \subset [p]$ there exists $l \in J \setminus [p]$ such that $\text{de}(l) \subseteq J$. We denote the topological ordering $(k_1, \ldots, k_p)$ and set $i = \max\{j \in [p] \colon k_j \notin J\}$. By definition, $l = k_i$ can only have descendants in J.



Q 10

Q 11 Wright's path analysis assumes all vertices are standardized. Thus we now have $\text{var}(A) = \text{var}(X) = \text{var}(Y) = 1$.

First, consider the paths starting from A to Y (or conveniently refer to the lecture notes), we have:

$$A \to M \to Y$$
$$A \leftarrow X \to Y$$
$$A \longleftrightarrow X \to Y$$
$$A \leftarrow X \longleftrightarrow Y$$

Second, consider the trek rule:

$$
\begin{aligned}
&A \longleftrightarrow A \to M \to Y \\
&A \leftarrow X \longleftrightarrow X \to A \to M \to Y \\
&A \longleftrightarrow X \to A \to M \to Y \\
&A \leftarrow X \longleftrightarrow A \to M \to Y \\
&A \leftarrow X \longleftrightarrow X \to Y \\
&A \longleftrightarrow X \to Y \\
&A \leftarrow X \longleftrightarrow Y
\end{aligned}
$$

The top four treks correspond to the first path (using σ algebra defined on the lecture notes they would give us the same value), and the fifth trek corresponds to the second path. The rest are the same. Thus the trek rule and Wright's path analysis to find $\mathrm{Cov}(A, Y)$ yield the same solution.

Q 12 First, we can write out the expression of the partial correlation between V_1 and V_2 given $\mathbf{V}_3$:

$$
\begin{aligned}
\mathrm{corr}(V_1, V_2|\mathbf{V}_3) &= \mathrm{corr}(V_1 - \Sigma_{13}\Sigma_{33}^{-1}\mathbf{V}_3, V_2 - \Sigma_{23}\Sigma_{33}^{-1}\mathbf{V}_3) \\
&= \frac{\Sigma_{12} - \Sigma_{13}\Sigma_{33}^{-1}\Sigma_{32}}{\sqrt{\Sigma_{11} - \Sigma_{13}\Sigma_{33}^{-1}\Sigma_{31}}\sqrt{\Sigma_{22} - \Sigma_{23}\Sigma_{33}^{-1}\Sigma_{32}}}
\end{aligned}
$$

Now we work on the inverse of matrix Σ, first treat the

$$
\Sigma = \left(\begin{array}{cc|c} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} \\ \Sigma_{21} & \Sigma_{22} & \Sigma_{23} \\ \hline \Sigma_{31} & \Sigma_{32} & \Sigma_{33} \end{array}\right) = \begin{pmatrix} E & F \\ G & H \end{pmatrix}
$$

Thus by using the formula for block matrix inversion, we have:

$$
\Omega = \Sigma^{-1} = \begin{pmatrix} S^{-1} & -S^{-1}FH^{-1} \\ -H^{-1}GS^{-1} & H^{-1} + H^{-1}GS^{-1}FH^{-1} \end{pmatrix} = \left(\begin{array}{cc|c} \Omega_{11} & \Omega_{12} & \Omega_{13} \\ \Omega_{21} & \Omega_{22} & \Omega_{23} \\ \hline \Omega_{31} & \Omega_{32} & \Omega_{33} \end{array}\right)
$$

Here only the upper left matrix inverse is of interest because it has the following elements:

$$
\begin{aligned}
S &= E - FH^{-1}G \\
&= \begin{pmatrix} \Sigma_{11} - \Sigma_{13}\Sigma_{33}^{-1}\Sigma_{31} & \Sigma_{12} - \Sigma_{13}\Sigma_{33}^{-1}\Sigma_{32} \\ \Sigma_{21} - \Sigma_{23}\Sigma_{33}^{-1}\Sigma_{31} & \Sigma_{22} - \Sigma_{23}\Sigma_{33}^{-1}\Sigma_{32} \end{pmatrix}
\end{aligned}
$$

Now we are dealing with a 2×2 matrix. Recall the previous expression of the partial correlation, we can pinpoint a few terms from the matrix S. We again use shorthand notation for the elements in S:

$$
S = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \quad \mathrm{corr}(V_1, V_2|\mathbf{V}_3) = b/\sqrt{ad}
$$

Move forward, we just need to inverse S:

$$
S^{-1} = \frac{1}{ad - bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix} = \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix}
$$

And it's obvious to show that:

$$
-\frac{\Omega_{12}}{\sqrt{\Omega_{11}\Omega_{22}}} = \frac{b}{\sqrt{ad}} = \mathrm{corr}(V_1, V_2|\mathbf{V}_3)
$$

Neyman (1923) considered the following simple difference-in-means estimator:

$$\hat{\beta}_{\mathrm{DIM}} = \bar{Y}_1 - \bar{Y}_0 \text{ where } \bar{Y}_1 = \frac{\sum_{i=1}^n A_i Y_i}{\sum_{i=1}^n A_i}, \bar{Y}_0 = \frac{\sum_{i=1}^n (1-A_i) Y_i}{\sum_{i=1}^n 1 - A_i}. \tag{2.5}$$

For the rest of this section, we will abbreviate $\hat{\beta}_{\mathrm{DIM}}$ as $\hat{\beta}$. Neyman studied the conditional distribution of $\hat{\beta}$ given the potential outcomes schedule. We refer to this as the *randomization distribution* of $\hat{\beta}$, because the only randomness here comes from randomizing the treatment $\boldsymbol{A}$. Of course, the randomization distribution depends on the randomization scheme. Neyman considered sampling $\boldsymbol{A}$ without replacement.

Let us first introduce some additional notation. Let the sample mean and variance of the potential outcome be

$$\bar{Y}(a) = \frac{1}{n}\sum_{i=1}^n Y_i(a),\ S^2(a) = \frac{1}{n-1}\sum_{i=1}^n \{Y_i(a) - \bar{Y}(a)\}^2,\ a = 0, 1.$$

Further, let the sample variance of the individual treatment effects be

$$S^2(0,1) = \frac{1}{n-1}\sum_{i=1}^n \{Y_i(1) - Y_i(0) - \beta_n\}^2.$$

Neyman's main result is summarized in the next theorem.

Theorem 2.2.1 *Consider the Neyman-Rubin causal model and suppose $\boldsymbol{A}$ is randomized by sampling without replacement with the constraint that $\sum_{i=1}^n A_i = n_1$ (Example 2.1.3). Let $\boldsymbol{W} = (Y_i(a))_{i\in[n],a\in\{0,1\}}$ be the potential outcomes schedule. Then the mean and variance of the randomization distribution of $\hat{\beta}$ are given by*

$$\mathbb{E}(\hat{\beta} \mid \boldsymbol{W}) = \frac{1}{n}\sum_{i=1}^n Y_i(1) - Y_i(0) = \beta_n, \tag{2.6}$$

$$\mathrm{Var}\left(\hat{\beta} \mid \boldsymbol{W}\right) = \frac{1}{n_0}S^2(0) + \frac{1}{n_1}S^2(1) - \frac{S^2(0,1)}{n}, \tag{2.7}$$

where $n_0 = n - n_1$.

Proof First, by using the consistency assumption, we may rewrite the difference-in-means estimator as

$$\hat{\beta} = \frac{1}{n_1}\sum_{i=1}^n A_i Y_i - \frac{1}{n_0}\sum_{i=1}^n (1-A_i) Y_i = \frac{1}{n_1}\sum_{i=1}^n A_i Y_i(1) - \frac{1}{n_0}\sum_{i=1}^n (1-A_i) Y_i(0). \tag{2.8}$$

By the linearity of expectation and $\mathbb{E}(A_i) = n_1/n$, the mean of $\hat{\beta}$ is given by

$$\begin{aligned}\mathbb{E}(\hat{\beta} \mid \boldsymbol{W}) &= \frac{1}{n_1}\sum_{i=1}^{n}\mathbb{E}(A_i)Y_i(1) - \frac{1}{n_0}\sum_{i=1}^{n}\mathbb{E}(1-A_i)Y_i(0) \\ &= \frac{1}{n_1}\sum_{i=1}^{n}\frac{n_1}{n}Y_i(1) - \frac{1}{n_0}\sum_{i=1}^{n}\frac{n_0}{n}Y_i(0) \\ &= \frac{1}{n}\sum_{i=1}^{n} Y_i(1) - Y_i(0).\end{aligned}$$

To prove (2.7), we follow the strategy outlined in Cox (2009, sec. 3.2).[5] Equation (2.8) shows that $\hat{\beta}$ is a linear function of $\boldsymbol{W}$, so $\mathrm{Var}\left(\hat{\beta} \mid \boldsymbol{W}\right)$ is a quadratic function of $\boldsymbol{W}$. Because the randomization scheme and the estimator $\hat{\beta}$ are invariant to permuting the units, the variance of $\hat{\beta}$ given $\boldsymbol{W}$ must also be permutation-invariant. Furthermore, in the special case that all the units have the same potential outcomes (i.e. $Y_1(0) = \cdots = Y_n(0)$ and $Y_1(1) = \cdots = Y_n(1)$), we have $\mathrm{Var}(\hat{\beta} \mid \boldsymbol{W}) = 0$. There, the randomization variance can be written as

$$\mathrm{Var}\left(\hat{\beta} \mid \boldsymbol{W}\right) = c_0 S^2(0) + c_1 S^2(1) + c_{01} S^2(0,1), \tag{2.9}$$

where c_0, c_1, and c_{01} are constants to be determined. Now consider the following bivariate normal model:

$$(Y_i(0), Y_i(1)) \sim \mathrm{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma_0^2 & \rho\sigma_0\sigma_1 \\ \rho\sigma_0\sigma_1 & \sigma_1^2\end{pmatrix}\right) \text{ independently, } i = 1, \ldots, n.$$

By the law of total variation, (2.6), and (2.9), we have

$$\begin{aligned}\mathrm{Var}(\hat{\beta}) &= \mathbb{E}(\mathrm{Var}(\hat{\beta} \mid \boldsymbol{W})) + \mathrm{Var}(\mathbb{E}(\hat{\beta} \mid \boldsymbol{W})) \\ &= \left[c_0\,\mathbb{E}(S^2(0)) + c_1\,\mathbb{E}(S^2(1)) + c_{01}\,\mathbb{E}(S^2(0,1))\right] + \mathrm{Var}(\beta_n) \\ &= c_0\,\mathrm{Var}(Y_1(0)) + c_1\,\mathrm{Var}(Y_1(1)) + \left(c_{01} + \frac{1}{n}\right)\mathrm{Var}\left(Y_1(1) - Y_1(0)\right) \\ &= c_0\sigma_0^2 + c_1\sigma_1^2 + \left(c_{01} + \frac{1}{n}\right)(\sigma_0^2 + \sigma_1^2 - 2\rho\sigma_0\sigma_1).\end{aligned}$$

On the other hand, consider first conditioning on $\boldsymbol{A}$. By noticing that $\mathbb{E}(\hat{\beta} \mid \boldsymbol{A}) = 0$ under the bivariate normal model, we have

$$\begin{aligned}\mathrm{Var}(\hat{\beta}) &= \mathbb{E}(\mathrm{Var}(\hat{\beta} \mid \boldsymbol{A})) \\ &= \mathbb{E}\Big(\sum_{i=1}^{n}\frac{1}{n_1^2}A_i^2\sigma_1^2 + \frac{1}{n_0^2}(1-A_i)^2\sigma_0^2 - \frac{2}{n_0 n_1}A_i(1-A_i)\rho\sigma_0\sigma_1\Big) \\ &= \mathbb{E}\Big(\sum_{i=1}^{n}\frac{1}{n_1^2}A_i\sigma_1^2 + \frac{1}{n_0^2}(1-A_i)\sigma_0^2\Big) \\ &= \frac{1}{n_1}\sigma_1^2 + \frac{1}{n_0}\sigma_0^2,\end{aligned}$$

[5] Cox (2009) assumed constant treatment effect (so $S^2(0,1) = 0$) but considered the more general randomized block design.

where the third equality uses $A_i \in \{0,1\}$. Equating the two expressions of $\mathrm{Var}(\hat{\beta})$ shows that $c_0 = 1/n_0$, $c_1 = 1/n_1$, and $c_{01} = -1/n$, which concludes our proof. □

Remark 2.2.2 To prove (2.7), a simpler but perhaps less illuminating approach involves directly computing the covariance matrix of $\boldsymbol{A}$; see Exercise 2.4. Our proof above is less straightforward but explains why the constants for S_0^2 and S_1^2 in (2.7) match those obtained from the normal theory (see also Imbens and Rubin, 2015, chap. 6, app. B) and can be more easily extended to other randomization schemes (Exercise 2.6).

The fact that potential outcomes appear on the right hand side of (2.6) and (2.7) should not come as a surprise, as the randomization distribution conditions on the potential outcomes schedule $\boldsymbol{W}$. This allows us to conclude from (2.6) that $\hat{\beta}$ is *unbiased* for β_n regardless of what the potential outcomes are. On the other hand, estimating the variance of $\hat{\beta}$ is difficult because the right-hand side of (2.7) depends on unknown potential outcomes. In particular, there is no hope to estimate $S^2(0,1)$ because we can never observe any individual treatment effect $Y_i(1) - Y_i(0)$ due to the fundamental problem of causal inference.

Fortunately, we can get a conservative estimator of $\mathrm{Var}\left(\hat{\beta} \mid \boldsymbol{W}\right)$ by pretending that $S_{01}^2 = 0$, i.e. there is no variation in individual treatment effects. The first two terms in (2.7) can be estimated by the sample variance of the observed outcomes within each treatment group

$$\hat{S}_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n} A_i (Y_i - \bar{Y}_1)^2, \;\; \hat{S}_0^2 = \frac{1}{n_0 - 1} \sum_{i=1}^{n} (1 - A_i)(Y_i - \bar{Y}_0)^2. \tag{2.10}$$

It is straightforward to verify that $\mathbb{E}(\hat{S}_a^2 \mid \boldsymbol{W}) = S^2(a)$, $a = 0, 1$ (Exercise 2.5). Putting these together, we obtain an variance estimator

$$\hat{V} = \hat{S}_0^2/n_0 + \hat{S}_1^2/n_1$$

that is conservative in the sense that $\mathbb{E}(\hat{V} \mid \boldsymbol{W}) \leq \mathrm{Var}\left(\hat{\beta} \mid \boldsymbol{W}\right)$. Under additional assumptions that restrict the variability of the potential outcomes, one can establish a central limit theorem for $\hat{\beta}$ (Li and Ding, 2017, thm. 5): for all "nice" $\boldsymbol{W}$,

$$\frac{\hat{\beta} - \beta_n}{\sqrt{\mathrm{Var}\left(\hat{\beta} \mid \boldsymbol{W}\right)}} \to \mathrm{N}(0,1) \text{ in distribution as } n \to \infty. \tag{2.11}$$

This can then be used to construct asymptotic tests and confidence intervals of β_n.

Note that in the central limit theorem in (2.11), the potential outcomes schedule $\boldsymbol{W} = (Y_i(a))_{i \in [n], a \in \{0,1\}}$ is fixed and the only random quantity on the left hand side is $\hat{\beta}$ (through its dependence on the treatment $\boldsymbol{A}$). In other words, equation (2.11) says $\hat{\beta}$ is asymptotically normal as long as the potential outcomes are not too variable. So even though the sample size n increases in infinity, equation (2.11) is inherently a *finite-population* statement. In contrast, the *super-population* (or repeated sampling) theory of statistical inference assumes that the