Conservative estimation of Difference-in-Means estimator's variance under random assignment*

Thomas Leavitt

The expression for the variance of the Difference-in-Means estimator, $\text{Var}_\Omega[\hat{\tau}]$, is

$$\text{Var}_\Omega[\hat{\tau}] = \frac{1}{n-1}\left(\frac{n_C\sigma_n^2(\boldsymbol{y_T})}{n_T} + \frac{n_T\sigma_n^2(\boldsymbol{y_C})}{n_C} + 2\sigma_n(\boldsymbol{y_C}, \boldsymbol{y_T})\right), \tag{1}$$

where

$$\sigma_n^2(\boldsymbol{y_T}) = \left(\frac{1}{n}\right)\sum_{i=1}^{n}\left(y_{Ti} - \frac{1}{n}\sum_{i=1}^{n} y_{Ti}\right)^2 \tag{2}$$

$$\sigma_n^2(\boldsymbol{y_C}) = \left(\frac{1}{n}\right)\sum_{i=1}^{n}\left(y_{Ci} - \frac{1}{n}\sum_{i=1}^{n} y_{Ci}\right)^2 \tag{3}$$

$$\sigma_n(\boldsymbol{y_C}, \boldsymbol{y_T}) = \left(\frac{1}{n}\right)\sum_{i=1}^{n}\left(y_{Ci} - \frac{1}{n}\sum_{i=1}^{n} y_{Ci}\right)\left(y_{Ti} - \frac{1}{n}\sum_{i=1}^{n} y_{Ti}\right). \tag{4}$$

Absent additional assumptions, we cannot unbiasedly estimate $\text{Var}_\Omega[\hat{\tau}]$. We can unbiasedly estimate the first two terms of $\text{Var}_\Omega[\hat{\tau}]$ in Equation (1), $\sigma_n^2(\boldsymbol{y_T})$ and $\sigma_n^2(\boldsymbol{y_C})$, but not the third term, $\sigma_n(\boldsymbol{y_C}, \boldsymbol{y_T})$, because no two potential outcomes of the same unit are observable. To conservatively estimate $\text{Var}_\Omega[\hat{\tau}]$, we will derive a quantity we can unbiasedly estimate that is always at least as great as the true variance, $\text{Var}_\Omega[\hat{\tau}]$.

*This is a live document that is subject to updating at any time.

The Cauchy-Schwarz inequality implies that

$$\sigma_n(\boldsymbol{y_C}, \boldsymbol{y_T}) \leq \sqrt{\sigma_n^2(\boldsymbol{y_C})\sigma_n^2(\boldsymbol{y_T})}$$

and the AM-GM inequality further implies that

$$\sqrt{\sigma_n^2(\boldsymbol{y_C})\sigma_n^2(\boldsymbol{y_T})} \leq \frac{\sigma_n^2(\boldsymbol{y_C}) + \sigma_n^2(\boldsymbol{y_T})}{2}.$$

Hence, it follows that $2\sigma_n(\boldsymbol{y_C}, \boldsymbol{y_T}) \leq \sigma_n^2(\boldsymbol{y_C}) + \sigma_n^2(\boldsymbol{y_T})$. Substituting $\sigma_n^2(\boldsymbol{y_C}) + \sigma_n^2(\boldsymbol{y_T})$ for $2\sigma_n(\boldsymbol{y_C}, \boldsymbol{y_T})$ therefore yields a tight upper bound for the true variance of the Difference-in-Means estimator:

$$\begin{aligned}
\mathrm{Var}_\Omega\left[\hat{\tau}\right] &= \frac{1}{n-1}\left(\frac{n_T\sigma_n^2(\boldsymbol{y_C})}{n_C} + \frac{n_C\sigma_n^2(\boldsymbol{y_T})}{n_T} + 2\sigma_n(\boldsymbol{y_C}, \boldsymbol{y_T})\right) \\
&\leq \frac{1}{n-1}\left(\frac{n_T\sigma_n^2(\boldsymbol{y_C})}{n_C} + \frac{n_C\sigma_n^2(\boldsymbol{y_T})}{n_T} + \sigma_n^2(\boldsymbol{y_C}) + \sigma_n^2(\boldsymbol{y_T})\right) \\
&= \frac{1}{n-1}\left(\frac{n_T\sigma_n^2(\boldsymbol{y_C})}{n_C} + \frac{n\sigma_n^2(\boldsymbol{y_T}) - n_T\sigma_n^2(\boldsymbol{y_T})}{n_T} + \sigma_n^2(\boldsymbol{y_C}) + \sigma_n^2(\boldsymbol{y_T})\right) \\
&= \frac{1}{n-1}\left(\frac{n_T\sigma_n^2(\boldsymbol{y_C})}{n_C} + \frac{n\sigma_n^2(\boldsymbol{y_T})}{n_T} - \frac{n_T\sigma_n^2(\boldsymbol{y_T})}{n_T} + \sigma_n^2(\boldsymbol{y_C}) + \sigma_n^2(\boldsymbol{y_T})\right) \\
&= \frac{1}{n-1}\left(\frac{n_T\sigma_n^2(\boldsymbol{y_C})}{n_C} + \frac{n\sigma_n^2(\boldsymbol{y_T})}{n_T} - \sigma_n^2(\boldsymbol{y_T}) + \sigma_n^2(\boldsymbol{y_C}) + \sigma_n^2(\boldsymbol{y_T})\right) \\
&= \frac{1}{n-1}\left(\frac{n_T\sigma_n^2(\boldsymbol{y_C})}{n_C} + \frac{n\sigma_n^2(\boldsymbol{y_T})}{n_T} + \sigma_n^2(\boldsymbol{y_C})\right) \\
&= \frac{1}{n-1}\left(\frac{n_T\sigma_n^2(\boldsymbol{y_C})}{n_C} + \frac{n_C\sigma_n^2(\boldsymbol{y_C})}{n_C} + \frac{n\sigma_n^2(\boldsymbol{y_T})}{n_T}\right) \\
&= \frac{1}{n-1}\left(\frac{n_T\sigma_n^2(\boldsymbol{y_C}) + n_C\sigma_n^2(\boldsymbol{y_C})}{n_C} + \frac{n\sigma_n^2(\boldsymbol{y_T})}{n_T}\right) \\
&= \frac{1}{n-1}\left(\frac{n_T\sigma_n^2(\boldsymbol{y_C}) + n\sigma_n^2(\boldsymbol{y_C}) - n_T\sigma_n^2(\boldsymbol{y_C})}{n_C} + \frac{n\sigma_n^2(\boldsymbol{y_T})}{n_T}\right) \\
&= \frac{1}{n-1}\left(\frac{n\sigma_n^2(\boldsymbol{y_C})}{n_C} + \frac{n\sigma_n^2(\boldsymbol{y_T})}{n_T}\right) \\
&= \frac{n}{n-1}\left(\frac{\sigma_n^2(\boldsymbol{y_C})}{n_C} + \frac{\sigma_n^2(\boldsymbol{y_T})}{n_T}\right).
\end{aligned}$$

We can unbiasedly estimate the two unknown parameters $\sigma_n^2(\boldsymbol{y_C})$ and $\sigma_n^2(\boldsymbol{y_T})$. Following Cochran

(1977, Theorem 2.4), unbiased estimators of $\sigma_n^2(\boldsymbol{y_C})$ and $\sigma_n^2(\boldsymbol{y_T})$ are

$$\widehat{\sigma}_n^2(\boldsymbol{y_C}) = \left(\frac{n-1}{n\,(n_C-1)}\right)\sum_{i=1}^{n}(1-Z_i)\left(y_{Ci}-\frac{1}{n_C}\sum_{i=1}^{n}(1-Z_i)Y_i\right)^2$$
$$\widehat{\sigma}_n^2(\boldsymbol{y_T}) = \left(\frac{n-1}{n\,(n_T-1)}\right)\sum_{i=1}^{n}Z_i\left(y_{Ti}-\frac{1}{n_T}\sum_{i=1}^{n}Z_iY_i\right)^2.$$

Therefore, the conservative estimator of the variance of the Difference-in-Means estimator is

$$\widehat{\mathrm{Var}_\Omega}\left[\hat{\tau}\right] = \frac{n}{n-1}\left(\frac{\widehat{\sigma}_n^2(\boldsymbol{y_T})}{n_T}+\frac{\widehat{\sigma}_n^2(\boldsymbol{y_C})}{n_C}\right) \tag{5}$$

This estimator is conservative in that $\mathrm{E}_\Omega\left[\widehat{\mathrm{Var}_\Omega}\left[\hat{\tau}\right]\right] \geq \mathrm{Var}_\Omega\left[\hat{\tau}\right]$. However, the estimator is unbiased when individual causal effects are constant, which implies that $2\sigma_n(\boldsymbol{y_C},\boldsymbol{y_T}) = \sigma_n^2(\boldsymbol{y_C})+\sigma_n^2(\boldsymbol{y_T})$.

We can also see that the conservative variance estimator is unbiased when individual causal effects are constant by examining the alternative expression for the Difference-in-Means estimator's variance given by

$$\mathrm{Var}_\Omega\left[\hat{\tau}\right] = \frac{S_n^2(\boldsymbol{y_T})}{n_T}+\frac{S_n^2(\boldsymbol{y_C})}{n_C}-\frac{S_n^2(\boldsymbol{\tau})}{n}, \tag{6}$$

where

$$S_n^2(\boldsymbol{y_T}) = \left(\frac{1}{n-1}\right)\sum_{i=1}^{n}\left(y_{Ti}-\frac{1}{n}\sum_{i=1}^{n}y_{Ti}\right)^2 \tag{7}$$

$$S_n^2(\boldsymbol{y_C}) = \left(\frac{1}{n-1}\right)\sum_{i=1}^{n}\left(y_{Ci}-\frac{1}{n}\sum_{i=1}^{n}y_{Ci}\right)^2 \tag{8}$$

$$S_n^2(\boldsymbol{\tau}) = \left(\frac{1}{n-1}\right)\sum_{i=1}^{n}\left(\tau_i-\frac{1}{n}\sum_{i=1}^{n}\tau_i\right)^2. \tag{9}$$

The conservative variance estimator equivalent to Equation (5) is

$$\widehat{\mathrm{Var}_\Omega}\left[\hat{\tau}\right] = \frac{1}{n_T}\widehat{S}_n^2(\boldsymbol{y_T})+\frac{1}{n_C}\widehat{S}_n^2(\boldsymbol{y_C}), \tag{10}$$

where

$$\widehat{S}_n^2(\boldsymbol{y_T}) = \left(\frac{1}{n_T-1}\right)\sum_{i=1}^{n}Z_i\left(y_{Ti}-\frac{1}{n_T}\sum_{i=1}^{n}Z_iY_i\right)^2$$

$$\widehat{S}_n^2(\boldsymbol{y_C}) = \left(\frac{1}{n_C - 1}\right) \sum_{i=1}^{n} (1 - Z_i) \left(y_{Ci} - \frac{1}{n_C} \sum_{i=1}^{n} (1 - Z_i) Y_i\right)^2,$$

assuming that $n_T \geq 2$ and $n_C \geq 2$.

This estimator in Equation (10) is conservative in that $\mathrm{E}_\Omega\left[\widehat{\mathrm{Var}}_\Omega\left[\hat{\tau}\right]\right] \geq \mathrm{Var}_\Omega\left[\hat{\tau}\right]$. However, this estimator of the Difference-in-Means estimator's variance is unbiased when individual causal effects are constant, i.e., $S_n^2(\boldsymbol{\tau}) = 0$.

References

Cochran, W. G. (1977). *Sampling Techniques* (3rd ed.). Hoboken, NJ: John Wiley & Sons. 2