Fundamentals

Core Terms

Fundamental Problem of Causal Inference: Only one potential outcome for each unit can be realized and thus observed.
Stable Unit Treatment Value Assumption:
– *No Interference*: The potential outcomes for any unit don't vary with the treatments of other units.
– *No Variations of Treatments*: The different versions of treatment levels are the same across units.

Paradoxes

Simpson's Paradox: Simpson's paradox occurs when a trend or relationship appears in different groups of data but disappears or reverses when these groups are combined. It highlights the danger of making conclusions based on aggregated data without accounting for the underlying factors that might be driving the observed trends.

For example, consider a situation where a treatment appears to have a positive effect when examining each subgroup separately, but when the subgroups are combined, the treatment effect appears negative. This can happen due to the presence of a confounding variable that interacts differently with the treatment in different subgroups. One such case is the below table:

Condition	Treatment	No Treatment
Severe	30/100 (30%)	125/250 (50%)
Mild	160/200 (80%)	45/50 (90%)
Total	190/300 (63%)	170/300 (57%)

Lord's Paradox: Lord's paradox is a counterintuitive phenomenon where controlling for a variable that is a common cause of both the predictor and the outcome can lead to unexpected changes in the relationship between the predictor and the outcome.

Here's an example: A large university is interested in investigating the effects on the students of the diet provided in the university dining halls and any sex differences in these effects. Various types of data are gathered. In particular, the weight of each student at the time of their arrival in September and their weight the following June are recorded. The results of the hypothetical study find that for males the average weight is identical at the end of the school year to what it was at the beginning; in fact, the whole distribution of weights is unchanged, although some males lost weight and some males gained weight — the gains and losses exactly balance. The same is true for females.

Statistician 1 observes that there are no differences between the September and June weight distributions for either males or females. Thus, Statistician 1 concludes that as far as these data are concerned, there is no evidence of any differential effect on the two sexes.

Statistician 2 observes that after "controlling for" initial weight, the diet has a differential positive effect on males relative to females because for males and females with the same initial weight, on average the males gain more than the females.

Experiments

Assignment Mechanisms

Individualistic Assignment: The following two conditions must apply:

- $P(W_i = 1|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = q(X_i, y_i(0), Y_i(1))$ for all units $i = 1, \ldots, N$ for some function $q(\cdot)$. This condition tells us that the unit assignment probability is independent and only depends on the covariates and potential outcomes of the unit itself (and not on the covariates or potential outcomes of other units).
- $P(\mathbf{W}|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = c \cdot \prod_{i=1}^{N} q(X_i, y_i(0), Y_i(1))^{W_i}(1 - q(X_i, y_i(0), Y_i(1)))^{1-W_i}$.
  This condition tells us that the probability of observing a specific treatment assignment vector $\mathbf{W}$ given the observed data and potential outcomes is proportional to the product of the probabilities of each unit being assigned treatment or control based on their potential outcomes.

Probabilistic Assignment: We can call an assignment mechanism probabilistic if the probability of assignment to treatment for each unit i is strictly between zero and one:

$$0 < P(W_i = 1|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) < 1, \text{ for } i = 1, \ldots, N$$

Unconfounded Assignment: We can call an assignment mechanism unconfounded if it doesn't depend on the potential outcomes:

$$P(W_i = 1|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = P(W_i = 1|\mathbf{X}, \mathbf{Y}'(0), \mathbf{Y}'(1))$$

Types of Experiments

A *randomized experiment* is an assignment mechanism that is (1) probabilistic and (2) has a known form controlled by the researcher.
A *classical randomized experiment* is a randomized experiment with an assignment mechanism that is (1) individualistic and (2) unconfounded.
The combination of individualistic and unconfounded assignment simplifies the assignment mechanism formula drastically to:

$$P(\mathbf{W}|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = c \cdot \prod_{i=1}^{N} e(X_i)^{W_i}(1 - e(X_i))^{1-W_i}$$

Here the assignment mechanism is just the product of propensity scores, which in turn change from being just the *average* assignment probability for units to just being the unit assignment probability.

Classical Randomzied Experiments

(1) Bernoulli Trial: Let the treatment probability for each unit i be some value $q(X_i)$ depending on the unit's covariates. So, each unit's assignment probability is a Bernoulli trial $\text{Bern}(e(X_i))$.

$$\text{Assignment Mechanism: } P(\mathbf{W}|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = c \cdot \prod_{i=1}^{N} e(X_i)^{W_i}(1 - e(X_i))^{1-W_i}$$

A common disadvantage of this setup is that there is a small probability (sometimes practically zero, but never exactly zero) that all units are assigned to the same treatment group.
(2) Completely Randomized Experiment: Decide on a fixed number of units that will be assigned to the treatment group, and randomly choose that many units to assign to that group.

$$\text{Assignment Mechanism: } P(\mathbf{W}|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = \begin{cases} \binom{N}{N_t}^{-1} & \text{if } \sum_{i=1}^{N} W_i = N_t \\ 0 & \text{otherwise.} \end{cases}$$

In this experiment type, all units have the same propensity score, namely N_t/N.
(3) Stratified Randomized Experiment: Divide the population into subgroups of J strata, so that the units within each group are similar with respect to some covariates we think are predictive of potential outcomes. Within each stratum conduct a completely randomized experiment.

$$\text{Assignment Mechanism: } P(\mathbf{W}|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = \begin{cases} \prod_{j=1}^{J} \binom{N(j)}{N_t(j)}^{-1} & \text{if the treatment setup is valid.} \\ 0 & \text{otherwise.} \end{cases}$$

Similar to the completely randomized setup, all units have the same propensity score $N_t(j)/N(j)$.
(4) Paired Randomized Experiment: An extreme version of the stratified randomized experiment design, where each stratum has exactly two units, and one of the two is assigned to the treatment group and the other to the control (with a probability of 1/2). The pairs are usually created based on a similarity ranking using covariates.

$$\text{Assignment Mechanism: } P(\mathbf{W}|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = \begin{cases} 2^{-N/2} & \text{if the treatment setup is valid.} \\ 0 & \text{otherwise.} \end{cases}$$

Fisher's Exact P-Value Approach

Overview

(1) Selecting a Null Hypothesis: A common choice of null hypothesis for this type of setup is a sharp null hypothesis – where we know the values of $\{\mathbf{Y}(0), \mathbf{Y}(1)\}$ if it is true. The most common choice of null is that of no effect for the active treatment:

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

Other nulls that fit under the sharp null paradigm include:

$$\text{Constant: } Y_i(1) - Y_i(0) = C, \text{ Logarithmic: } Y_i(1)/Y_i(0) = C, \text{ for all } i$$

As such, under the sharp null, both potential outcomes are known by either direct observation or inferred under the hypothesis used.
(2) Select a Test Statistic A test statistic is a real-valued function of the treatment assignments, the observed outcomes, and the pre-treatment covariates $(\mathbf{W}, \mathbf{Y}^{\text{obs}}, \mathbf{X})$. There are multiple ways to reason which test statistic should be used. Most importantly, however, is understanding that test statistics are considered stochastic only through the stochastic nature of the assignment vector, and as such acquire "randomization distributions". Here are the three methods of selecting a test statistic, with associated motivations:

1. **Simple Additive Effect**: Although not necessarily the best test statistic to use, a natural and popular choice is the absolute value of the difference in average outcomes by treatment group, which works well in cases of a presumed additive treatment effect:

$$T^{\text{dif}} = \left|\bar{Y}_t^{\text{obs}} - \bar{Y}_c^{\text{obs}}\right| = \left|\frac{\sum_{i:W_i=1} Y_i^{\text{obs}}}{N_t} - \frac{\sum_{i:W_i=0} Y_i^{\text{obs}}}{N_c}\right|$$

2. **Unbalanced Groups**: In cases where the sizes of groups being unbalanced is a concern, using a t-statistic based statistic would be useful:

$$T^{t-\text{statistic}} = \left|\frac{\bar{Y}_t^{\text{obs}} - \bar{Y}_c^{\text{obs}}}{\sqrt{s_c^2/N_c + s_t^2/N_t}}\right|, \text{ where the } s^2 \text{ values are the sample variances.}$$

3. **Multiplicative Effect**: In cases where we presume the effect to be multiplicative (and using a log-based sharp null), a log-transformed statistic would be useful:

$$T^{\log} = \left|\frac{\sum_{i:W_i=1} \ln(Y_i^{\text{obs}})}{N_t} - \frac{\sum_{i:W_i=0} \ln(Y_i^{\text{obs}})}{N_c}\right|$$

   Such a statistic would make sense if raw data have skewed distributions – typically the case for positive variables like earnings, wealth, or levels of a pathogen, and as such treatment effects are more likely to be multiplicative than additive. However, using this statistic would also require adjusting values if any observed outcome is 0, as the logarithm of 0 is undefined.

4. **Outlier Robustness**: In cases where we have observed outcomes with some considerable outliers, statistics that are robust to outliers can come in handy. Some possible options include:
   - *Median Based*: $T^{\text{median}} = \left|\text{median}_t(Y_i^{\text{obs}}) - \text{median}_c(Y_i^{\text{obs}})\right|$
   - *Quantile Based*: $T^{\text{quantile}} = \left|Q_t(Y_i^{\text{obs}}) - Q_c(Y_i^{\text{obs}})\right|$, where Q is the quantile function of the empirical distribution of observed outcomes.
   - *Rank Based*: $T^{\text{rank}} = |\bar{R}_t - \bar{R}_c|$, where R is the rank of a particular observation in the general pool of observed outcomes. It is sometimes useful to normalize ranks.

5. **Statistical/Linear Models**: If we would like to build models that describe relationships within the data, and we have parameters in these models that are different for each of the control and treatment groups, then the difference in those parameters can be used (this motivates the use of regression in estimation).

6. **Covariate Use**: Covariates can be used in creating statistics, especially if a covariate can be interpreted as an observation of the response variable at a time before the study took place:
   - *Additively*: One way to do so is by considering *gain scores*, defined as:

$$Y_i'(w) = Y_i(w) - X_i$$

     This motivates the following test statistic:

$$T^{\text{gain}} = \frac{\sum_{i:W_i=1} Y_i'^{\text{obs}}}{N_t} - \frac{\sum_{i:W_i=0} Y_i'^{\text{obs}}}{N_c} = \bar{Y}_t^{\text{obs}} - \bar{Y}_c^{\text{obs}} - (\bar{X}_t - \bar{X}_c)$$

   - *Proportionally*: Another way is to consider the proportional change from a baseline value:

$$T^{\text{proportion-change}} = \bar{Y_t}'' - \bar{Y_c}'' = \frac{1}{N_t}\sum_{i:W_i=1} \frac{Y_i^{\text{obs}} - X_i}{X_i} - \frac{1}{N_c}\sum_{i:W_i=0} \frac{Y_i^{\text{obs}} - X_i}{X_i}$$

     Both the gain score and the proportional change from baseline statistics are likely to lead to more powerful tests if the covariate X_i is a good proxy for $Y_i(0)$. This often happens when a covariate is a lagged version of the outcome variables, like a test-score or an income value.

(3) p-value Calculation: Can be done either exactly or using simulation, depending on whether the number of possible treatment assignments under the null is small (equal to $\binom{N}{N_t}$). If not, we randomly draw treatment vectors and calculate statistics to generate an empirical distribution of the test statistic. *Note:* Multiple comparison problems require adjusting the α level.

(4) Fisher Intervals: We can repeat this process for an array of different hypothesized individual treatment effects τ_{H_0} and see whether that would result in rejecting the null or not. The set of values for which we do not reject their associated null, as a result, can be thought of as a sort of confidence interval (not really a confidence interval, but a "Fisher interval").

Statistical Power

The statistical power of a test is the probability of rejecting the null hypothesis given that the alternative hypothesis is true. We can also calculate the statistical power of a Fisher's Exact Test using a specific null and alternative hypotheses:

$$H_0 : Y_i(1) - Y_i(0) = \gamma, \text{ for all } i = 1, \ldots, N, \text{ and } H_1 : Y_i(1) - Y_i(0) = \delta, \text{ for all } i = 1, \ldots, N$$

We do so using the following formula:

$$\sum_{\mathbf{w}',\mathbf{y}'} 1\left\{\left[\sum_{\mathbf{w},\mathbf{y}} 1\{T(\mathbf{w},\mathbf{y}) \geq T(\mathbf{w}',\mathbf{y}')\}\frac{1\{\mathbf{w}\in\mathcal{W}^+, (\mathbf{w},\mathbf{y})\in\mathcal{Y}^+_{\tau=\gamma}\}}{\binom{N}{N_t}}\right] \leq \alpha\right\}\frac{1\{\mathbf{w}'\subset\mathcal{W}^+, (\mathbf{w}',\mathbf{y}')\in\mathcal{Y}^+_{\tau=\delta}\}}{\binom{N}{N_t}}$$

Neyman's Repeated Sampling Approach

Neyman's treatment effect examines the average of the unit level treatment effect:

$$\tau_{fs} = \overline{Y(1)} - \overline{Y(0)} = \frac{1}{N}\sum_{i=1}^{N}(Y_i(1) - Y_i(0))$$

We usually use the following *unbiased* estimator for it:

$$\hat{\tau}^{\text{dif}} = \bar{Y}_t^{\text{obs}} - \bar{Y}_c^{\text{obs}} = \frac{\sum_{i:W_i=1} Y_i^{\text{obs}}}{N_t} - \frac{\sum_{i:W_i=0} Y_i^{\text{obs}}}{N_c}$$

The variance of this estimator can be expressed in two ways:

$$\text{Var}(\bar{Y}_t^{\text{obs}} - \bar{Y}_c^{\text{obs}}) = \frac{s_c^2}{N_c} + \frac{s_t^2}{N_t} - \frac{s_{tc}^2}{N}, \text{ or equivalently } = \frac{N_t \cdot s_c^2}{N \cdot N_c} + \frac{N_c \cdot s_t^2}{N \cdot N_t} + \frac{2}{N}\rho_{tc} \cdot s_c \cdot s_t$$

In the first formula, the first two terms are within-group variances and the third represents treatment effect heterogeneity. With greater heterogeneity, individual treatment effects vary more widely, leading to a larger spread of values and higher variance of the estimator. However, the presence of both positive and negative treatment effects across individuals can lead to some degree of cancellation when averaging across the sample; individuals with positive treatment effects may compensate for ones with negative treatment effects to an extent, resulting in a reduction of variance in the estimator.

An unbiased (if τ is constant for all units), but conservative estimator for the variance is:

$$\hat{\mathbf{V}}^{\text{neyman}} = \frac{s_c^2}{N_c} + \frac{s_t^2}{N_t} = \frac{1}{N_t - 1}\sum_{i:W_i=1}(Y_i(1) - \bar{Y}_t^{\text{obs}})^2 + \frac{1}{N_c - 1}\sum_{i:W_i=0}(Y_i(0) - \bar{Y}_c^{\text{obs}})^2$$

Confidence intervals for the average treatment effect usually rely on an appeal to the Central Limit Theorem, where the $1-\alpha\%$ confidence interval is:

$$\text{CI}^{1-\alpha}_{\text{neyman}} = (\hat{\tau}^{\text{dif}} - z_{\alpha/2} \cdot \sqrt{\hat{\mathbf{V}}^{\text{neyman}}}, \hat{\tau}^{\text{dif}} + z_{1-\alpha/2} \cdot \sqrt{\hat{\mathbf{V}}^{\text{neyman}}})$$

Super Population Inference

Under the super-population point of view, we consider the potential outcome vectors $\mathbf{Y}(1), \mathbf{Y}(0)$ to be a random sample of size N from the super-population. We expect super-population inference to be less precise than the finite sample setting, as variance now exists due to two sources: variance due to the treatment assignment $\mathbf{W}$ and variance due to the sampling of units from the super-population. The estimate in this kind of inference is slightly different:

$$\tau_{sp} = \text{E}(Y_i(1) - Y_i(0))$$

With the same estimator as in the finite sample setting:

$$\hat{\tau}^{\text{dif}} = \bar{Y}_t^{\text{obs}} - \bar{Y}_c^{\text{obs}} = \frac{\sum_{i:W_i=1} Y_i^{\text{obs}}}{N_t} - \frac{\sum_{i:W_i=0} Y_i^{\text{obs}}}{N_c}$$

The variance can be derived as follows, by using Eve's law and remembering that $\mathbf{Y}(1), \mathbf{Y}(0)$ are now random as well:

$$\begin{aligned}\text{Var}(\bar{Y}_t^{\text{obs}} - \bar{Y}_c^{\text{obs}}) &= \text{E}(\text{Var}(\bar{Y}_t^{\text{obs}} - \bar{Y}_c^{\text{obs}}|\mathbf{Y}(1), \mathbf{Y}(0))) + \text{Var}(\text{E}(\bar{Y}_t^{\text{obs}} - \bar{Y}_c^{\text{obs}}|\mathbf{Y}(1), \mathbf{Y}(0))) \\ &= \text{E}\left(\frac{s_c^2}{N_c} + \frac{s_t^2}{N_t} - \frac{s_{tc}^2}{N}\right) + \text{Var}(\bar{Y}(1) - \bar{Y}(1)) = \frac{\sigma_c^2}{N_c} + \frac{\sigma_t^2}{N_t}\end{aligned}$$

For which the Neyman variance estimator is an unbiased (and not conservative) estimator.

Regression Analysis

Useful Terminology:

Finite Sample Average Treatment Effect	τ_{fs}	$\frac{1}{N}\sum_{i=1}^{N}(Y_i(1)-Y)i(0))$
Super Population Average Treatment Effect	τ_{sp}	$\mathrm{E}(Y_i(1)-Y_i(0))$
S. P. Average of $Y_i(0)$ Conditional on Covariates	$\mu_c(x)$	$\mathrm{E}(Y_i(0)\|X_i=x)$
S. P. Average of $Y_i(1)$ Conditional on Covariates	$\mu_t(x)$	$\mathrm{E}(Y_i(1)\|X_i=x)$
S. P. Variance of $Y_i(0)$ Conditional on Covariates	$\sigma_c^2(x)$	$\mathrm{Var}(Y_i(0)\|X_i=x)$
S. P. Variance of $Y_i(1)$ Conditional on Covariates	$\sigma_t^2(x)$	$\mathrm{Var}(Y_i(1)\|X_i=x)$
Mean of Unit Level Treatment Effect Given Covariates	$\tau(x)$	$\mathrm{E}(Y_i(1)-Y_i(0)\|X_i=x)$
Variance of Unit Level Treatment Effect Given Covariates	$\sigma_{ct}^2(x)$	$\mathrm{Var}(Y_i(1)-Y_i(0)\|X_i=x)$

We can also use standard regression analysis for causal inference. In fact, a linear regression setup without the use of any covariates is similar to previously encountered methods:

$$Y_i^{\text{obs}} = \alpha + \tau \cdot W_i + \epsilon_i$$

Here, we can solve for the unknown coefficient using the following equations:

$$\hat{\tau}^{\text{OLS}} = \frac{\sum_{i=1}^{N}(W_i-\bar{W})\cdot(Y_i^{\text{obs}}-\bar{Y}^{\text{obs}})}{\sum_{i=1}^{N}(W_i-\bar{W})^2}$$

$$\hat{\alpha}^{\text{OLS}} = \bar{Y}^{\text{obs}} - \hat{\tau}^{\text{OLS}}\cdot\bar{W}$$

In this setting, $\hat{\tau}^{\text{OLS}}$ is identical to the difference in average outcomes by treatment status – what we've been investigating using usual methods! One minor difference between usual methods and regression methods is the meaning of the estimands:

$$\tau = \mathrm{E}(Y_i(1)|W_i=1) - \mathrm{E}(Y_i(0)|W_i=0)$$

$$\alpha = \mathrm{E}(Y_i(0)|W_i=0)$$

- The estimand τ is the mean of the outcome under exposure among those who are exposed minus the mean of the outcome under no exposure among those who are not exposed.
- The estimand α is the mean of the outcome under no exposure among those who are not exposed.

The OLS estimators for these estimands are consistent (unbiased for large samples) for both the case where we do or don't use pre-treatment covariates. The consistency of the least squares estimator for τ_{sp} is not affected by the accuracy of the regression function specification in a completely randomized experiment. Regardless of how nonlinear the conditional expectations of potential outcomes given covariates are in the entire population, simple least squares regression reliably estimates the population average treatment effect. This is because, through random treatment assignment, the correlation between the treatment indicator and the covariates in the entire population is eliminated. While this correlation might deviate from zero in finite samples, it diminishes in large samples, rendering the inclusion of covariates insignificant for estimation purposes.

Another important feature to inspect is the variance of our estimator $\hat{\tau}$. There are two ways to calculate it, depending on whether homoskedasticity holds:

1. Homoskedastic Variance Estimator: Assumes equal variances for all observations.

$$\left(\frac{1}{N_c}+\frac{1}{N_t}\right)\frac{1}{N-2}\sum_{i=1}^{N}\hat{\epsilon}_i^2$$

2. Robust Sandwich Estimator: A heteroskedastic variance estimator.

$$\frac{\sum_{i=1}^{N}\hat{\epsilon}_i^2(W_i-\bar{W})^2}{\left(\sum_{i=1}^{N}(W_i-\bar{W})^2\right)^2} = \frac{((N_t-1)/N_t)s_t^2}{N_t} + \frac{((N_c-1)/N_c)s_c^2}{N_c}$$

We can use an appeal to the Central Limit Theorem to create an asymptotic distribution for $\hat{\tau}$, and as such create confidence intervals for it using our chosen variance estimator from above.

Observational Studies

To tackle observational studies, we first depart from the conventional assumption of a known function governing treatment assignment, characteristic of randomized experiments. As before, we continue to uphold the *unconfoundedness* principle, which asserts that assignment is independent of potential outcomes. Additionally, we maintain an *individualistic* assignment mechanism, where the likelihood of treatment for a unit relies solely on its pre-treatment variables, and a *probabilistic* one, ensuring that the chance of receiving treatment falls between zero and one for all units.

These assumptions suggest that the assignment mechanism can be likened to conducting a completely randomized experiment within subpopulations of units sharing the same covariate values. In fact, if we have a limited number of categorical covariates using which we can divide the dataset into a small number of strata, we can run usual analyses without the need to explicity know the propensity scores/treatment probabilities. This, however, is somewhat tricky for continuous pre-treatment covariates, so new methods are necessary.

In most if not all observational study approaches, balancing score are a key concept. A balancing score is a function denoted as $b(x)$, where x represents the covariates. This function is designed to satisfy the property that when you condition on it, the treatment assignment W_i becomes independent from the covariates X_i. In simpler terms, a balancing score is a way to create groups or conditions where the treatment assignment is not influenced by the covariates. It helps ensure that when comparing different groups or conditions in a study, any differences observed are more likely to be due to the treatment itself rather than other factors related to the covariates. The propensity score is the most popular balancing score, but also the coarsest one; the propensity score is a function of every possible balancing score.

When dealing with many pre-treatment covariates (including continuous ones), there are two groups of strategies when dealing with observational data, specifically when our end goal is to estimate τ_{sp}:

- Imputation Based Approach: Use a linear regression model to impute the missing potential outcomes from the observational data, then run analysis as usual. We then estimate the treatment effect using the following expression:

$$\hat{\tau}^{\text{OLS}} = \frac{1}{N}\sum_{i=1}^{N}\left(W_i\cdot(Y_i^{\text{obs}} - X_i\hat{\beta}_c^{\text{OLS}}) + (1-W_i)\cdot(X_i\hat{\beta}_t^{\text{OLS}} - Y_i^{\text{obs}})\right)$$

  Generally, we tend to fit two linear regression models, one for each camp of the observed potential outcomes. This method is not recommended in cases where the covariate distributions are different between the treated and control groups.

- Propensity Scoring Approach: Approximate the propensity scores in some way; we will call these estimated propensity scores $\hat{e}(X_i)$. Then, use one of the following three methods to estimate $\hat{\tau}_{sp}$:

  1. Estimator Weighting: Weigh each observation by its weight/propensity score relative to other propensity scores to get an estimate for τ_{sp}:

$$\hat{\tau}^{\text{HT}} = \frac{1}{N}\sum_{i:W_i=1}\hat{\lambda}_i Y_i^{\text{obs}} - \frac{1}{N}\sum_{i:W_i=0}\hat{\lambda}_i Y_i^{\text{obs}},\ \text{where: } \hat{\lambda}_i = \begin{cases} N\cdot\frac{(1-\hat{e}(X_i))^{-1}}{\sum_{i:W_i=0}(1-\hat{e}(X_i))^{-1}} & \text{if } W_i=0 \\ N\cdot\frac{\hat{e}(X_i)^{-1}}{\sum_{i:W_i=1}\hat{e}(X_i)^{-1}} & \text{if } W_i=1 \end{cases}$$

     This method is also not recommended in cases where the covariate distributions are different between the treated and control groups.

  2. Blocking Estimator/Subclassification: A more robust approach involving the propensity score is to coarsen it through blocking/subclassification. We partition the sample into subclasses, based on the value of the estimated propensity score. Within each subclass, the data can be analyzed as if they arose from a completely randomized experiment. Notationally, let $b_j, j=0,1,...,J$ denote the subclass boundaries, with $b_0=0$ and $b_J=1$, and let $B_i(j)$ be a binary indicator, equal to 1 if $b_{j-1}<\hat{e}(X_i)<b_j$, and zero otherwise. To estimate the within group treatment effect we use the following expression:

$$\hat{\tau}_{\text{dif}(j)} = \frac{\sum_{i:B_i(j)=1}Y_i\cdot W_i}{\sum_{i:B_i(j)=1}W_i} - \frac{\sum_{i:B_i(j)=1}Y_i\cdot(1-W_i)}{\sum_{i:B_i(j)=1}(1-W_i)}$$

     Then, to estimate the overall finite-sample average effect of the treatment, we use the following expression:

$$\hat{\tau}_{\text{strat}} = \sum_{j=1}^{J}\frac{N(j)}{N}\cdot\hat{\tau}_{\text{dif}(j)}$$

While this approach offers greater robustness compared to the weighting estimator when dealing with units having extreme values of the estimated propensity score, we advise against using it without certain modifications. Specifically, to mitigate bias and enhance precision, it is useful to implement covariance adjustment within the subclasses.

3. Matching Estimators: In contrast to model-based imputation, weighting, and blocking techniques, matching doesn't consistently require estimating an unknown function. Rather, it hinges on identifying direct comparisons, or matches, for each unit. When considering a treated unit with specific covariate values, the objective is to find a control unit with a closely matching set of covariates.

Assessing covariate balance in observational studies is crucial because it helps ensure that the groups being compared are similar in terms of covariates, aside from the treatment or exposure of interest. This validates the use of methods such as regression for imputation and estimator weighting. This can either be done by comparing descriptive statistics of the covariate distribution across treatment and control, or by comparing the distributions of estimated propensity scores.

Estimating Propensity Scores

We usually estimate propensity scores using logistic regression. We do this by using a stepwise selection method for which covariates to include in the model:

1. **Initial K_B Features**: Initially, we opt to incorporate K_B fundamental covariates based on their substantive relevance. These covariates may encompass factors deemed crucial for elucidating the assignment process and potentially linked to certain outcome measures. Alternatively, KB may equal zero if the researcher possesses limited substantive insight into the covariates' comparative significance.
2. **Additional First-Order Features (K_L Features Total)**: In the second phase, we choose certain covariates from the remaining pool for integration into the specification of the propensity score; we denote the number of these remaining covariates as $K - K_B$. Dduring the covariate selection process, given we have already chosen K_L linear terms, including the K_B terms chosen in the initial step. We face the decision of whether to incorporate an additional covariate from the set of $K - \tilde{K}_L$ remaining covariates, and if so, which one. This decision-making process relies on the outcomes of $K - \tilde{K}_L$ additional logistic regression models. In each of these models, we compute the likelihood ratio statistic, evaluating the null hypothesis that the newly included covariate possesses a coefficient of zero. If all the likelihood ratio statistics fall below a predefined constant C_L, we stop our step-wise inclusion process. However, if at least one likelihood ratio test statistic exceeds CL, we include the covariate associated with the largest likelihood ratio statistic.
3. **Second-Order Features (K_Q Features Total)**: In the last phase, we simply repeat the second step, deciding which second order terms (squared features and interaction terms) to include in our models. We do this analgously, with a different (oftentimes higher) cutoff for the likelihood ratio statistic, C_Q.

After choosing a model, and to asses the propensity scoring estimator, we need to define strata boundaries such that within each stratum, the propensity varies minimally. We create those groupings (i.e. select the boundaries) using the following procedure:

1. **Drop Extreme Units**: Initially, we exclude from analysis all control units with an estimated propensity score lower than the smallest value among the treated units' estimated propensity scores, i.e. $\min \hat{e}(X_i)$, where $W_i = 1$, as well as all treated units with an estimated propensity score higher than the largest value among the control units' estimated propensity scores, i.e. $\min \hat{e}(X_i)$, where $W_i = 0$. This trimming procedure guarantees a degree of overlap between the two groups.
2. **Assess Block Adequacy**: In this procedure, we utilize the estimated linearized propensity score (or log odds ratio), denoted as $\hat{e}(x) = \ln\left(\frac{\hat{e}(x)}{1-\hat{e}(x)}\right)$. The rationale behind focusing on the linearized propensity score instead of the propensity score itself lies in its tendency to exhibit a distribution that can be more accurately approximated by a normal distribution. Utilizing the linearized propensity scores, we assess the following two conditions for each block $j = 1, \ldots, J$:
   - **Independence**: Is the estimated linearized propensity score within the block approximately uncorrelated with the treatment indicator? We evaluate this using a t-statistic. Let $N_c(j)$ and $N_t(j)$ represent the subsample sizes for controls and treated units in block j, calculated as $N_{c(j)} = \sum_{i=1}^{N}(1 - W_i) \cdot B_i(j)$ and $N_{t(j)} = \sum_{i=1}^{N} W_i \cdot B_i(j)$, respectively. Let $\bar{c}(j)$ and $\bar{t}(j)$ denote the average values for the estimated linearized propensity score, categorized by treatment status and block, given by:

$$\bar{c}(j) = \frac{1}{N_{c(j)}} \sum_{i=1}^{N}(1 - W_i) \cdot B_i(j) \cdot \hat{e}(X_i)$$

$$\bar{t}(j) = \frac{1}{N_{t(j)}} \sum_{i=1}^{N} W_i \cdot B_i(j) \cdot \hat{e}(X_i)$$

and let $S^2(j)$ denote the sample variance of the linearized propensity score within block j, calculated as:

$$S^2(j) = \frac{1}{N(j) - 2}\left(\sum_{i:B_i(j)=1}(1 - W_i) \cdot (\hat{e}(X_i) - \bar{c}(j))^2 + \sum_{i:B_i(j)=1} W_i \cdot (\hat{e}(X_i) - \bar{t}(j))^2\right)$$

The t-statistic for block j is then defined as:

$$t_j = \frac{\bar{t}(j) - \bar{c}(j)}{\sqrt{S^2(j) \cdot \left(\frac{1}{N_{c(j)}} + \frac{1}{N_{t(j)}}\right)}}$$

We compare this t-statistic for each stratum to a predetermined threshold value, denoted as $t_{\max}$, for instance, $t_{\max} = 1$.

- **Strata Splitting Criteria**: If we were to divide the current jth stratum into two sub-strata, what would be the new boundary value, and how many observations would fall into each of the new sub-strata? We determine the median value of the propensity score among the $N_{c(j)} + N_{t(j)}$ units in a particular block and denote this median as b_j. To be precise, if the current number of units in the stratum $N_{c(j)} + N_{t(j)}$ is odd, the median is the middle value, and if the number of units in the stratum is even, the median is defined as the average of the two middle values. Then, with the superscripts l and h denoting the low and high sub-stratum respectively, we calculate the sizes of the relative groups:

$$N^l_{c(j)} = \sum_{i=1}^{N}(1 - W_i) \cdot B_i(j) \cdot \mathbf{1}_{\{\hat{e}(X_i) < b_j\}}, \quad N^l_{t(j)} = \sum_{i=1}^{N} W_i \cdot B_i(j) \cdot \mathbf{1}_{\{\hat{e}(X_i) < b_j\}}$$

$$N^u_{c(j)} = \sum_{i=1}^{N}(1 - W_i) \cdot B_i(j) \cdot \mathbf{1}_{\{\hat{e}(X_i) \geq b_j\}}, \quad N^u_{t(j)} = \sum_{i=1}^{N} W_i \cdot B_i(j) \cdot \mathbf{1}_{\{\hat{e}(X_i) \geq b_j\}}$$

The current block j is considered inadequately balanced if the t-statistic is too high, $|t_j| > t_{\max}$, and amenable to splitting if the number of units in each new block of each treatment type is sufficiently large to allow for a split at the median, $\min(N^l_{c(j)}, N^l_{t(j)}, N^u_{c(j)}, N^u_{t(j)}) \geq 3$, and $\max(N^l_{c(j)}, N^l_{t(j)}, N^u_{c(j)}, N^u_{t(j)}) \geq K + 2$, where K is the number of pre-treatment variables. We choose these numbers so that we can compare mean covariate values within blocks and so that later we can make adjustments for any remaining covariate differences within blocks.

3. **Split Blocks Where Appropriate**: If a block is inadequately balanced and the sub-blocks that could be created have appropriate sizes, we split the block along the median propensity value within the block. This procedure is repeated until blocks cannot be split any further either due to size or to their being adequately balanced.

Next, with these strata defined, it would be great if we could assess the quality of our propensity score estimator. This can be done by (most simply) be comparing the means of the covariates across treatment and control for each stratum using a hypothesis test; we do need to note that this a repeated testing problem so the significance level has to be adjusted.
In actually estimating the treatment effect, we do the following: create the propensity scorer, create bounds, and then asses the quality of the scorer. If everything is working correctly, use linear regression to estimate the treatment effect within each stratum (adjusted for differences in covariates within each stratum). Finally, find the average treatment effect by summing over all strata in proportion to how large they are.