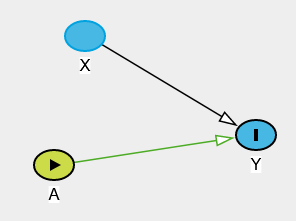
# Matching

## Observational Studies

### Randomised Trials

* Treatment assignment A is determined by a coin toss, which effectively erases the arrow from X to A
  + 
  + No back-door path from A to Y
* The distribution of X (covariates) is the same in both treatment groups as the original distribution.
* The distribution of pre-treatment variables X that affect Y are the same in both treatment groups – covariate balance.
* If the outcome distribution ends up differing, it’s not because the differences in X i.e. isolated the treatment effect.
* Covariates are dealt with at the design phase.

### Why Not Always Randomise?

* Expensive (much more than observational studies).
* Can be unethical.
* Many people refuse to participate – don’t want to be experimented on, don’t want to deal with the hassle of being followed up on etc.
* Take time (must wait for outcome data) – and then the question may not then be relevant.

### Observational Studies

* Two different types: planned and retrospective.
* Planned/prospective with active data collection:
  + Like trials: data collected on a common set of variables at planned times; outcomes carefully measured; study protocols.
  + Unlike trials: regulations much weaker since not intervening (just observing); broader population eligible for the study (fewer restrictions around who you can/can’t include).
* Databases, retrospective, passive data collection e.g. medical records; claims; registries
  + Data collected for a purpose other than research/experimentation.
  + Large sample sizes; inexpensive; potential for rapid analysis (data already sitting there).
  + Data quality typically lower; no uniform collection standard.
* Distribution of confounders/covariates differs between treatment groups.
  + For example, if older people are more likely to get treated then the distribution for that group may be sifted to the right relative to the group of younger people.

### Matching

* Method that attempts to make an observational study more like a randomised trial.
* Main idea: match subjects in the treatment and control groups on the set of important covariates, X.
  + For example, for any particular covariate value, there should be about the same number of treated and untreated subjects.

### Advantages

* Controlling for confounders achieved at the design phase – without looking at the outcome (which does exist but we ignore it).
  + i.e. the difficult statistical work is conducted completely blind to the outcomes.
  + The hard work involves finding good matches.
  + Similarly to randomising where we’re controlling for X by randomly assigning treatment without looking at the outcome (which doesn’t exist, but that’s beside the point).
* Matching reveals lack of overlap in covariate distribution.
  + Relative to other approaches for confounder control.
  + i.e. there could be some treated subjects who are unlike anyone in the control group, which suggests they essentially had no chance of not getting treated, which is useful because they should be excluded from the study.
  + Don’t want to make inferences about subjects who had 0% chance of being in the control group, as that would involve extrapolation.
  + Positivity assumption will hold in the population that can be matched.
    - Can identify violations of this assumption and exclude those subjects from the research.
    - If don’t match, may not notice a violation of this assumption.
* Once matches, data can essentially be treated as if they came from a randomised trial.
* Outcome analysis is therefore relatively simple.

## Overview of Matching

### Single Covariate

* Recall the goal is to match treated and control subjects on their covariates (the set identified as sufficient to control for confounding).
* Assume there’s a single covariate: hypertension.
  + Partially determines which treatment a subject is likely to get and hence is something we’d like to control for.
  + Without matching, there’s likely to be a big imbalance between the groups on this covariate in an observational study, whereas in a randomised trial, it’s going to be balanced.
* Match each treated subject to a control subject on the values on the covariate and ignore/exclude the rest of the possible controls.
  + End up with perfect balance on the covariate.

### Many Covariates

* More complicated.
* Not able to perfectly match on the full set of covariates.
* In a randomised trial, treated and control subjects won’t be perfectly matched either, but if the distribution of covariates is balanced between the groups, we achieve stochastic balance.
* Matching closely on covariates can achieve stochastic balance.
  + Ideally need one of the groups to be a lot bigger than the other to maximise the chances of finding good matches for as many of the smaller group as possible.
  + This will typically be the control group.

### Target Population

* Making the distribution of the covariates in the control population look like that in the treated population: Causal effect of treatment on the treated.
  + Making inferences about the treated population rather than the entire population.
* Can do matching where you make the treated and control groups look like each other and like the population as a whole, but this is more complicated (and beyond the scope of this course).

### Fine Balance

* Sometimes it’s difficult to find great matches.
* Fine balance: treated and control groups have the same covariate distribution, even if there are non-ideal matches (don’t have stochastic balance).
  + For example, average age and percent female are the same in both groups even though the matched pairs aren’t great.
    - Match 1: treated, male, age 40 and control, female, age 45.
    - Match 2: treated, female, age 45 and control, male, age 40.

### Number of Matches

* One-to-one (pair matching): match exactly one control to every treated subject.
* Many to one: match some fixed number K controls to every treated subject.
* Variable: match as many good matches as can be found for a single treated subject.
* Overall, one-to-one should result in better matches because you’re not trying to match as many control subjects to a given treated subject, but you’re discarding more data, so could decrease efficiency.

## Matching Directly on Confounders

### How to Match?

* Need to choose some metric of closeness. For example, two measures of distance are:
  + Mahalanobis distance
  + Robust Mahalanobis distance.

### Mahalanobis Distance

* The Mahalanobis distance (M-distance) between covariates for subject and subject is:
  + Square root of the sum of squared distances between each covariate scaled by the covariance matrix.
    - We need to scale because ‘big’ is relative i.e. the units of different covariates is different. End up with all variables on the same scale.
* Ultimately, use this metric to compare a given treated subject with all the controls to help locate a good match.
* For example, assume we have 3 confounders: age, COPD (1 = yes, 0 = no), female (1 = yes, 0 = no)

|  |  |  |
| --- | --- | --- |
| **Treated** | | |
| **Age** | **COPD** | **Female** |
| 78.17 | 0 | 1 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Control** | | | |
| **Age** | **COPD** | **Female** | **Distance** |
| 70.25 | 1 | 0 | 4.23 |
| 75.33 | 0 | 1 | 0.17 |
| 86.08 | 1 | 1 | 3.72 |
| 54.97 | 0 | 0 | 2.45 |
| 43.63 | 0 | 0 | 2.89 |
| 18.04 | 0 | 1 | 3.60 |

### Robust Mahalanobis Distance

* Outliers can create large distances between subjects, even if the covariates are otherwise similar.
* Ranks could be more relevant
  + Similar to classical statistics in which original values are replaced by ranks.
  + E.g. highest and second-highest ranked values of covariates perhaps should be treated as similar, even if the values are far apart.
* That is:
  + Replace each covariate value with its rank.
  + Constant diagonal on the covariance matrix – ranks should be all on the same scale and don’t want to weight one covariate more than another.
  + Calculate the M-distance on the ranks.

### Other Distance Measures

* Distance metrics are important because you need to know what ‘close’ means.
* Can do some clever things to get exact matches on some variables and tolerate inexact matches on others / weight some covariates more than others.
  + For example, if you want an exact match on a few important covariates, you can essentially make the distance infinity if they’re not equal to make it impossible for them to match.
* Distance match on the propensity score
  + Common approach.

## Greedy Nearest-Neighbour Matching

* Intuitive.
* Computationally fast.
  + Simple algorithms (e.g. identifying minimum distance).
  + Fast even for large datasets.
  + R package: MatchIt.
* Sensitive to data order i.e. randomisation/shuffling.
* Not optimal.
  + Always taking the smallest distance match doesn’t necessarily minimise the total distance.
  + Can lead to some bad matches.

### Set-Up

* Selected a set of pre-treatment covariates X that satisfy the ignorability assumption.
* Calculated a distance between each treated subject and every control subject.
* You have more controls than treated subjects.
* For now, we’re going to focus on one-to-one or pair matching.

### Greedy Matching

1. Randomly order the list of treated and control subjects.
   1. Not strictly necessary, but good practice in case the subjects are ordered in some way.
2. Start with the first treated subject. Match to the control with the smallest distance.
   1. Going for the best matching control, rather than considering the wider picture. This is why it’s considered greedy.
3. Remove the matched control from the pool of available controls.
4. Move on to the next treated subject and repeat until all treated subjects have a matched control.

### Example

* Assume we have 2 treated subjects and 6 controls.

|  |  |  |
| --- | --- | --- |
| **Treated** | | |
| **Age** | **COPD** | **Female** |
| 78.17 | 0 | 1 |
| 67.91 | 0 | 0 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Control** | | | | |
| **Age** | **COPD** | **Female** | **Distance 1** | **Distance 2** |
| 70.25 | 1 | 0 | 4.23 | 3.70 |
| 75.33 | 0 | 1 | 0.17 | ~~2.06~~ |
| 86.08 | 1 | 1 | 3.72 | 4.34 |
| 54.97 | 0 | 0 | 2.45 | 0.78 |
| 43.63 | 0 | 0 | 2.89 | 1.45 |
| 18.04 | 0 | 1 | 3.60 | 3.60 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Matched Dataset** | | | | |
| **Match ID** | **Treated** | **Age** | **COPD** | **Female** |
| 1 | 1 | 78.17 | 0 | 1 |
| 1 | 0 | 75.33 | 0 | 1 |
| 2 | 1 | 67.91 | 0 | 0 |
| 2 | 0 | 54.97 | 0 | 0 |

### Many-to-One Matching

* For k:1 matching
  + After all subjects have a single match, go back through the data and find the 2nd best match.
  + Repeat until there are k matches for each subject.

### Trade-Offs

* Pair matching
  + Closer matching in general – finding best match you can between the treated and control group.
  + Faster computing time – fewer matches.
* Many-to-one
  + Will get some bad matches.
  + Larger sample size – use more of the controls so more efficient estimates of the causal effects.
    - But efficiency gain isn’t quite as good as you might first think. Better to have an additional matched pair.
* Effectively a bias-variance trade-off issue.
  + Pair matching: less bias because the matching is closer but higher variance because you’re discarding data.
  + Many-to-one matching: higher bias but less variance because you’re matching more data.
  + Where do you need to be along the spectrum/What’s more important?

### Caliper

* Might prefer to exclude treated subjects for whom there doesn’t exist a good match.
* A bad match can be defined using a caliper – maximum acceptable distance.
  + Only accept a match to a treated subject if the best match has a distance less than the caliper.
  + Recall the positivity assumption: the probability of each treatment given X should be non-zero.
    - If there are no matches within the caliper, it suggests that the positivity assumption has been violated.
      * It indicates that there is really no one like them in the control group and hence they never had much of a chance of not being treated.
    - Excluding these subjects makes the assumption more realistic… BUT the population becomes harder to define.
      * i.e. the population is now all treated subjects for which a match was possible instead of just all treated subjects.
* Nevertheless, useful to use a caliper and not accept ‘bad’ matches.

## Optimal Matching

* Minimises the global distance measure e.g. total distance.
* Computationally demanding especially for larger datasets.
  + Essentially a network flow optimisation problem.
* R packages include optmatch and rcbalance.

### Greedy Matching Isn’t Optimal

* Doesn’t necessarily lead to the smallest total distance i.e. isn’t globally optimal.
* For example, suppose we want to match on age, and have 3 treated subjects and 12 control subjects.
* Greedy matching

|  |  |  |  |
| --- | --- | --- | --- |
| **Treated** | **Available Controls** | **Matched Treated** | **Matched Control** |
| 45  38  41 | 72 ~~47~~  ~~44~~ 54  60 ~~36~~  63 71  ~~35~~ 56  65 27 | 45  38  41 | 44  36  47 / 35 |

* + Total distance = 1 + 2 + 6 = 9
* Selected a better set of matches

|  |  |  |  |
| --- | --- | --- | --- |
| **Treated** | **Available Controls** | **Matched Treated** | **Matched Control** |
| 45  38  41 | 72 ~~47~~  ~~44~~ 54  60 ~~36~~  63 71  35 56  65 27 | 45  38  41 | 47  36  44 |

* + Total distance = 2 + 2 + 3 = 7
* May be times you wish to accept a slightly worse match to achieve a better match later on and hence achieve an overall better set of matches/smaller total distance between the treated and control group.
  + In greedy matching, this can never occur.

### Computational Feasibility

* Depends on the size of the problem, where size is defined as the number of possible treatment-control pairings.
  + For example, 100 treated subjects and 1,000 controls = 100,000 possible pairings.
  + 1 million treatment-control pairings is feasible on most PCs (but don’t expect an answer quickly).
  + 1 billion pairings (e.g. 10,000 treated subjects and 100,000 control subjects) may not be.

### Sparse Optimal Matching

* Constraints can be imposed to reduce the computational overhead for large datasets.
  + Doesn’t result in full global optimal matching, but a good compromise that reduces the size of the optimisation problem.
  + For example, matching within hospitals for a multi-site clinical study; matching within primary disease category.
  + These are known as ‘blocks’.
* Known as sparse matching and mismatches are tolerated if fine balance is still achievable.
  + Accept mismatches as long as the resulting distributions of the covariates in the treated and control groups are similar.

## Assessing Balance

### Did Matching Work?

* Need to assess whether it worked.
  + Covariate balance
    - Several ways of doing this – for example, standardised differences – similar means?
  + Should be done without looking at the outcome variable.
  + If the conclusion is no, then should go back and revisit the matching algorithm.
* Commonly, a ‘Table 1’ is created where pre-matching and post-matching balance is compared.
* Hypothesis tests i.e. test for difference in means between the treated and control group for each covariate.
  + Two-sample t-tests (continuous) /chi-squared (discrete/categorical) - report p-values for each test
  + P-values are dependent on sample size so small differences in means could still be significant for large sample sizes.

### Standardised Differences

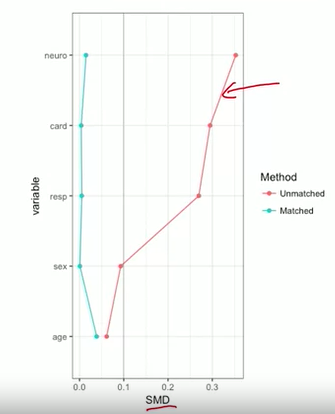
* Typically preferred to the hypothesis test approach as it’s not dependent on the sample size.
* Difference in means between groups, divided by the pooled standard deviation:
  + Difference in means in standard deviation units.
    - smd = 1 implies there’s a 1 sd difference in means, which is large.
    - Always relative to the standard deviation, as we don’t want the smd to change if we change scales.
* Often reported in absolute value terms.
* Calculated for each covariate.
* Rules of thumb:
  + Values < 0.1 indicate adequate balance.
  + Values 0.1 – 0.2 are acceptable.
  + Values > 0.2 indicate serious imbalance (and the matching algorithm should be revisited).

### Table 1

* Often standardised differences are displayed in a table before/after matching.
* Summary statistics of the main variables involved in the analysis.
* Typically stratify the table on the treatment groups if you’re carrying out an unmatched analysis or comparing treatments.
* For example:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Unmatched** | | |
|  | **No RHC** | **RHC** | **SMD** |
| n | 3551 | 2184 |  |
| age (mean (sd)) | 61.8 (17.3) | 60.8 (15.6) | 0.06 |
| sex = Yes (%) | 53.9 | 58.5 | 0.09 |
| resp = Yes (%) | 41.7 | 28.9 | 0.27 |
| card = Yes (%) | 28.4 | 42.3 | 0.3 |
| neuro = Yes (%) | 16.2 | 5.4 | 0.35 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Matched** | | |
|  | **No RHC** | **RHC** | **SMD** |
| n | 2082 | 2082 |  |
| age (mean (sd)) | 61.6 (16.7) | 61.0 (16.7) | 0.039 |
| sex = Yes (%) | 56.9 | 56.9 | 0.001 |
| resp = Yes (%) | 30.6 | 30.4 | 0.005 |
| card = Yes (%) | 39.3 | 39.5 | 0.004 |
| neuro = Yes (%) | 5.3 | 5.7 | 0.015 |

* 2nd table is what we hope for after matching i.e. the statistics look similar to what they would have had we conducted a RCT
* Can also plot the SMDs, which is particularly useful if there are many covariates.
  + Can really highlight how well the matching algorithm performed overall.
  + 
  + 0.1 is the general rule-of-thumb indicating a ‘good’ balance.

## Analysing (Outcome) Data After Matching

### After Matching

* Assume that the matching has been successful and there is adequate balance in the covariates.
* Outcome analysis may include:
  + Test for a treatment effect (hypothesis testing).
  + Estimate a treatment effect and confidence interval.
  + Methods should account for the matching process.

### Randomisation Tests

* Also known as permutation or exact tests.
* Used if you actually had conducted an RCT, but can act as though you had given a matching process was undertaken.
* Main idea:
  + Compute test statistic from observed data.
  + Assume the null hypothesis of no treatment effect is true.
  + Randomly permute the treatment assignment within pairs and re-compute the test statistic.
  + Repeat multiple times to determine how different the observed test statistic is from the permuted test statistics.
    - If the null hypothesis is actually true, the observed and permuted test statistics should be very similar.

### Example 1 – Categorical Data

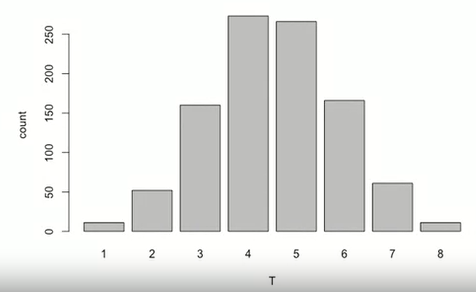
* Assume we have a binary outcome and 13 matched pairs.
* Test statistic: number of events (how many times the outcome occurred) in the treated group.
* Observed data

|  |  |  |
| --- | --- | --- |
| **Matched Pair ID** | **Treated** | **Control** |
| 1 | 0 | 0 |
| 2 | 1 | 0 |
| 3 | 1 | 0 |
| 4 | 0 | 0 |
| 5 | 1 | 1 |
| 6 | 0 | 0 |
| 7 | 0 | 1 |
| 8 | 0 | 1 |
| 9 | 0 | 0 |
| 10 | 1 | 0 |
| 11 | 0 | 0 |
| 12 | 1 | 0 |
| 13 | 1 | 0 |

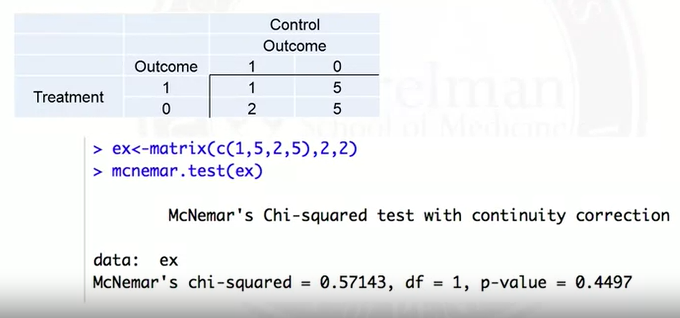
* Test statistic = 6
* Randomly permute the values in each row – note that this can only occur in discordant pairs as permutation won’t change anything in concordant pairs.
  + All the information about a treatment effect comes from discordant pairs.
  + Permutation means randomly re-labelling who is a control and who is a treatment subject in each pair.
  + Could perform this manually by drawing from a binomial distribution with probability 0.5 (in this case, could be different for a different outcome variable).
* Permutation 1

|  |  |  |
| --- | --- | --- |
| **Matched Pair ID** | **Treated** | **Control** |
| 1 | 0 | 0 |
| 2 | 1 | 0 |
| 3 | 0 | 1 |
| 4 | 0 | 0 |
| 5 | 1 | 1 |
| 6 | 0 | 0 |
| 7 | 1 | 0 |
| 8 | 0 | 1 |
| 9 | 0 | 0 |
| 10 | 0 | 1 |
| 11 | 0 | 0 |
| 12 | 1 | 0 |
| 13 | 1 | 0 |

* Test statistic = 5
* If we permute say 1,000 times and record the test statistics, we get the following distribution:

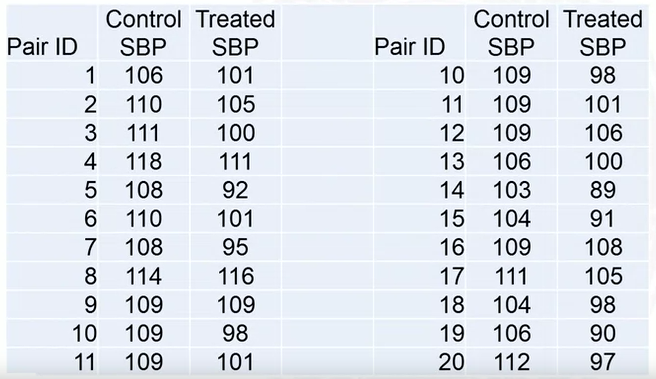


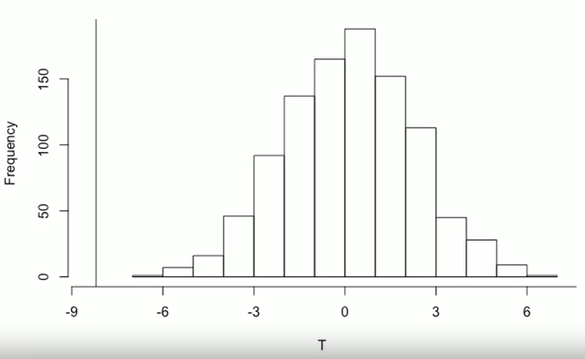
* The observed test statistic isn’t inconsistent with the null hypothesis so there is no evidence to conclude that the treatment was doing anything.
* To calculate a p-value, add up the probabilities of getting a value equal to or more extreme than the observed test statistic (here 6, 7, 8, 1, 2, 3 for a two-sided test).
* Equivalent to the McNemar test for paired data:



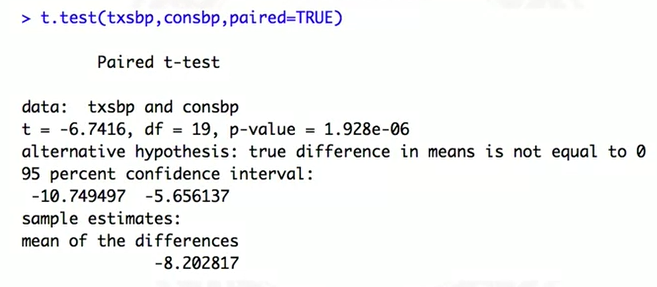
* + These are counts of treatment and control in the observed data.
  + All the information about whether there’s a treatment effect comes from the discordant pairs, which here is comparing the 5 and the 2.
  + If there’s no treatment effect, you’d expect the numbers in the off-diagonal to be about the same.
  + 0.4497 would be about the same as if you added up the counts in the columns 1, 2, 3, 6, 7, 8 and divided by 1000 (which again suggests that there’s no evidence of a treatment effect).

### Example 2 – Continuous Data

* The basic approach also works for continuous data.
* Assume you have blood pressure outcome data from 20 matched pairs.
* Test statistic: difference in sample means.
* 
* Test statistic = -8.2 mm Hg
  + Stratify the treatment group, calculate a sample mean for each and take the difference.
* Randomly reassign who is a treatment and who is a control in each pair, and repeat 1,000 times to calculate the test statistic distribution.



* Typically centred around 0 if there’s no treatment effect.
* The observed test statistic is way out in the left tail i.e. very unusual and the p-value is around 0.
* A paired t-test confirms that the null hypothesis should be rejected in this example (i.e. strong evidence of a treatment effect):



### Other Outcome Models

* Conditional logistic regression.
  + Matched binary outcome data.
* Stratified Cox model.
  + Time-to-event (survival) outcome data.
  + Baseline hazard stratified on matched sets.
* Generalised estimating equations (GEE)
  + Match ID variable used to specify clusters.
  + For binary outcomes, can estimate a causal risk difference, causal risk ratio, or causal odds ratio (depending on the link function).

## Sensitivity Analysis

### Possible Hidden Bias

* Matching aims to achieve balance on observed covariates – variables selected ahead of time believed to be sufficient to control for confounding.
  + Overt bias occurs if there is imbalance on these covariates.
* No guarantee matching will result in balance on variables not matched on incl. unobserved variables.
  + Different to RCTs where there should be balance on both observed and unobserved variables because treatment assignment is truly random.
  + Hidden bias occurs if these unobserved variables are confounders – ignorability assumption is violated.
    - Unmeasured confounding in the analysis.

### Sensitivity Analysis

* If there is hidden bias present, how severe would it have to be to change the conclusions of the study?
  + Change in statistical significance conclusion.
  + Change in the direction of the effect.
* Typically assume that there is always some hidden bias/unmeasured confounding in a study.
* Let and be the probability that subject and subject receives the treatment respectively.
* Suppose subjects and are perfectly matched, such that their observed covariates and are the same.
* There is no hidden bias if .
  + There are no hidden/unobserved variables that could be affecting the treatment assignment probability.
* Consider the following inequality:
  + Numerator is the odds of treatment for subject .
  + Denominator is the odds of treatment for subject .
  + is the odds ratio.
  + If , then there’s no overt bias.
  + If , then there’s hidden bias; in particular, subject is more likely to receive treatment than subject even though they have the same observed covariates.
* can therefore quantify the degree to which the assumption of no hidden bias is violated.
  + : assumption is fine.
  + (close to 1): assumption is also fine – barely violated.
  + (much larger than 1): assumption is violated.

### Example

* Suppose we have evidence of a treatment effect under the assumption of no hidden bias.
  + Assuming we have carried out the process described in early sections: i.e. matched control to treatment, concluded there is suitable balance in the covariates etc.
* Carry out sensitivity analysis:
  + Increase until evidence of a treatment effect disappears i.e. no longer statistically significant.
    - If this occurs when (for example), then the results are very sensitive to hidden bias/unmeasured confounding.
      * The probability of treatment for one subject doesn’t have to change very much for the conclusion is changed.
    - If this occurs when , then the results aren’t particularly sensitive to unmeasured confounders.
      * And hence can be more confident in the conclusions.
  + R packages that can perform sensitivity analysis include sensitivity2x2xk, sensitivityfull.

# Propensity Scores

## Propensity Scores

* Propensity score – probability of receiving treatment rather than the control, given the covariates X
* Propensity score for subject :
* For example:
  + Suppose age was the only X variable and older people were more likely to get treatment.
  + Propensity score would be larger for older ages:
* Interpretation: If subject has a propensity score value of 0.3, it means that there is a 30% change they will be treated given their covariate values.
* Typically, a propensity score won’t be unique given the covariate values i.e. there could be more than one distinct combination of covariates that results in the same propensity score.

### Balancing Score

* Suppose 2 subjects have the same value of the propensity score, but possibly different covariate values X.
  + Despite the different covariate values, they were both equally likely to be treated.
  + Hence, both subject’s covariate combinations were just as likely to be found in the treatment group.
  + If you restrict to a subpopulation of subjects with the same propensity score value, there should be balance in the two treatment groups.
  + So, the propensity score is a balancing score – i.e. if you condition on it then you’ll have balance.
    - If you restrict the sample to subjects with the same propensity score and stratify on actual treatment received, you should see the same covariate distribution in the two groups.
  + What is the distribution of covariates among treated subjects with a specific propensity score? This is the same as the distribution of covariates of control subjects with the same propensity score.
  + Implication: If we match on the propensity score, we should achieve balance (if done well).
    - This follows if we assume ignorability – treatment assignment is randomised given X.
    - Conditioning on the propensity score is conditioning on an allocation probability (the rate at which treatment should be assigned).

### Estimated Propensity Score

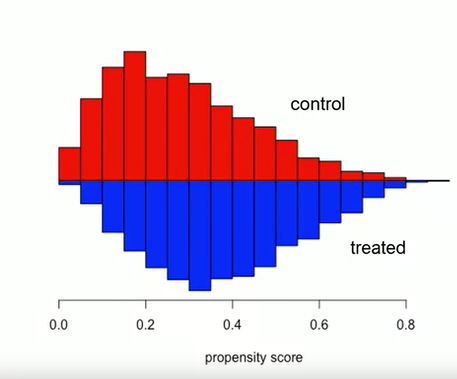
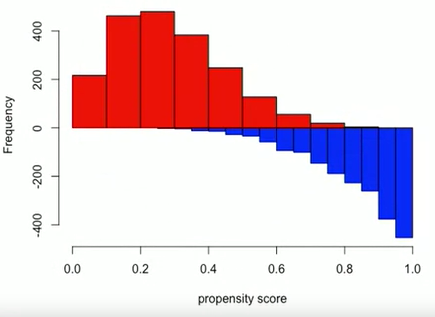
* In a randomised trial, the propensity score is generally known:
  + Researchers typically decide the allocation probability, which is generally 0.5 as it won’t depend on a subject’s characteristics.
  + Could have some sort of stratified sampling, but either way the allocation probability is known ahead of time and hence the propensity score is known ahead of time.
* In an observational study, it’s unknown as we don’t know ahead-of-time the probability of treatment given a subject’s covariates.
* But since the propensity score calculation involves observed data, it can be estimated.
* Typically, when a propensity score is referred to, it’s in the context of an estimated score, not actual.
* Need to estimate .
  + The ‘outcome’ here is A, which is binary, so propensity scores can be estimated using something like a logistic regression (or any classification algorithm).
  + Fit a model to the outcome based on all important covariates.
  + Then, get the predicted probability (fitted values) for each subject – the estimated propensity score.

## Propensity Score Matching

### Matching

* We previously noted that the propensity score is a balancing score, so matching on the propensity score should achieve balance even though we didn’t match directly on covariates.
  + i.e. balance on the covariate distribution between treated and control subjects.
* Propensity scores are scalars – each subject will have exactly one value. This means the matching problem is simplified to a single ‘summary’ variable – summarising all the covariates.

### Overlap

* Once propensity scores are estimated, it’s useful to look at the overlap in the propensity score distributions prior to matching i.e. comparing the treated and control distributions.
  + Often done with a plot.
  + Particularly interested in whether all subjects had at least a non-zero probability of receiving treatment.
* Example 1: 
  + Good overlap everywhere (some subjects in both groups irrespective of the estimated propensity score) – what you want to see if you want to conduct propensity score matching.
  + Positivity assumption appears reasonable.
  + Peak/center of the treated group is shifted slightly right relative to the control group, which is expected.
* Example 2: 
  + Poor overlap – some propensity scores only appearing in the treatment/control group.
  + Positivity assumption likely violated.
  + Region for which we could hope to learn about causal effects would be restricted to about 0.4 to 0.8 – regions for which there is overlap in propensity scores.
    - Can’t expect to learn about a treatment effect for subjects guaranteed to either be in the control group or the treatment group.
    - One solution could be to remove subjects with extreme propensity scores and just focus on the common region (region of common support).

### Trimming Tails

* If there is a lack of overlap, trimming the tails is an option i.e. remove subjects with extreme propensity score values.
  + For example, noting we expect the treatment group to have higher propensity scores on average.
    - Remove control subjects with propensity scores less than the minimum in the treatment group.
    - Remove treatment subjects with propensity scores greater than the maximum in the control group.
* Makes the positivity assumption more reasonable and prevents extrapolation.
  + If you tried to estimate a causal effect for subjects with very small propensity scores in the second example, you’d actually be extrapolating because you don’t have any information in that region on treated subjects i.e. there’s not enough information in the data about the relationship between the outcome and propensity scores.

### Matching

* Compute a distance between the propensity scores for every treated subject and every control.
  + Use the same types of matching algorithms as before.
* In practice, the logit transformation/log-odds of the propensity score is often used, rather than the propensity score itself.
  + Propensity scores are bounded between 0 and 1, making many values seem similar.
    - For example, if the treatment was rare then you could have the propensity scores bunched up in a small range.
  + The logit is unbounded, so the distribution is effectively stretched out whilst preserving the ranks/order of the propensity scores, which means that matches are easier to find.

### Caliper

* Prevents bad matches being accepted.
* The maximum distance tolerated between the propensity scores of the treated and control pair.
* In practice, it’s common to choose the caliper based on standard deviation units i.e. 0.2 times the standard deviation of the logit of the propensity score:
  + Estimate propensity scores.
  + Logit-transform the scores.
  + Calculate the standard deviation.
  + Multiply the standard deviation by 0.2.
* Hence, the caliper is based on the variability in the propensity scores, so a larger variability would imply a higher tolerance to differences, and a smaller variability a lower tolerance.
* Commonly done because it appears to work well (multiplying by 0.2), but it’s somewhat arbitrary and can be tuned based on the results of the matching when balance is assessed.
* Bias-variance trade-off where a smaller caliper means better balance but fewer matched pairs so the treatment effect estimate will have greater variability.

### After Matching

* Outcome analysis methods are exactly the same as if matching had been performed on the covariates.