Outline

1. Causal inference
   - Confounding and lurking variables
   - Imbalance and lack of overlap
   - Fundamental problem of causal inference
     - Blocking
     - Randomization
     - Statistical adjustment

Causal inference versus prediction

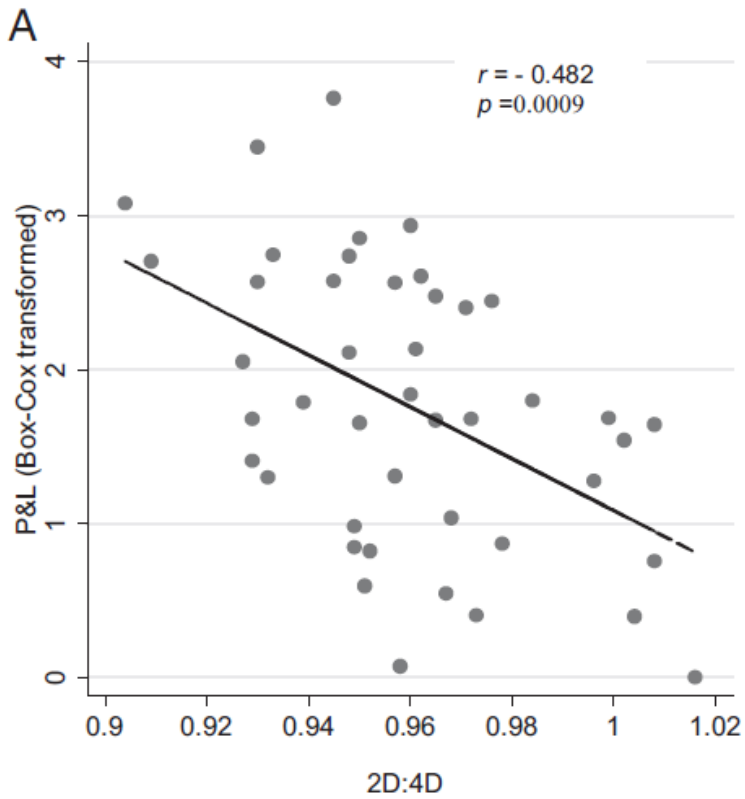
- In prediction, we make comparisons between outcomes across different combinations of values of input variables.
- In causal inference, we ask what would happen to an outcome *y* as a result of a treatment or intervention.
- Predictive inference relates to comparisons between units.
- Causal inference addresses comparisons of different treatments when applied to the same unit.

Example of prediction, not causal inference

The second-to-fourth digit (2D:4D) ratio (length of index finger / ring finger) was used to predict financial traders' performance (profit and loss: P&L). It was stated to serve as a surrogate for prenatal androgen effects. (Coates et al. 2009 PNAS)

Second-to-fourth digit ratio predicts success among high-frequency financial traders

John M. Coates[a,b,1], Mark Gurnell[c,1], and Aldo Rustichini[d,1]



Counterfactual outcomes

Let T represent a treatment variable.

- For a categorical treatment,

$$T_i = \mathbb{1}_i \text{ is treated} = \begin{cases} 1 & \text{if unit } i \text{ receives the treatment} \\ 0 & \text{if unit } i \text{ receives the control} \end{cases}$$

$y_i^1 =$ outcome of the ith unit if the treatment is given.
$y_i^0 =$ outcome of the ith unit if the control is given.

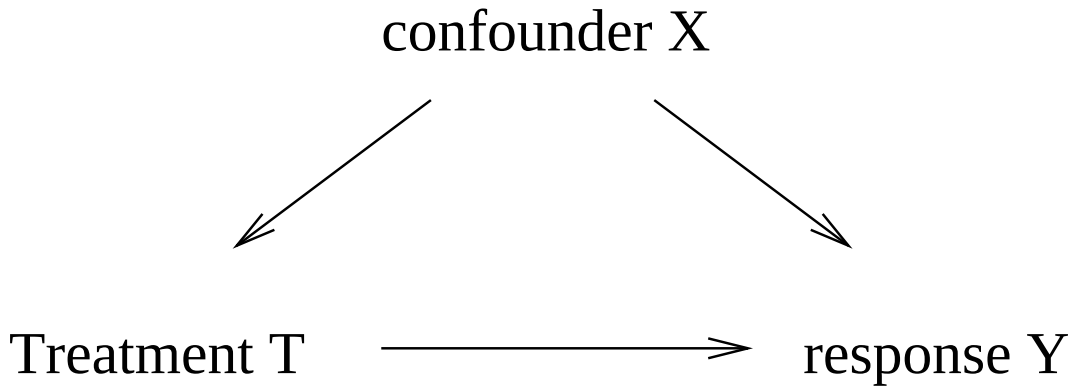
One of these is observed, the other is counterfactual — what would have been observed if the other treatment had been given?

- For a continuous treatment,

$T_i =$ the numerical value of the treatment assigned to unit i

Confounding

A variable (or covariate) is a confounder if it predicts both treatment and outcome.



We can estimate a causal effect in regression if:

1. the regression model includes all confounders; and
2. the regression model is correct

A confounder left out of a regression model is called a lurking variable.

Causal inference and estimation of treatment effects can be misleading when confounders are omitted from a model.

Ex: $Y =$ FEV, $T =$ smoking, and $X =$ age.

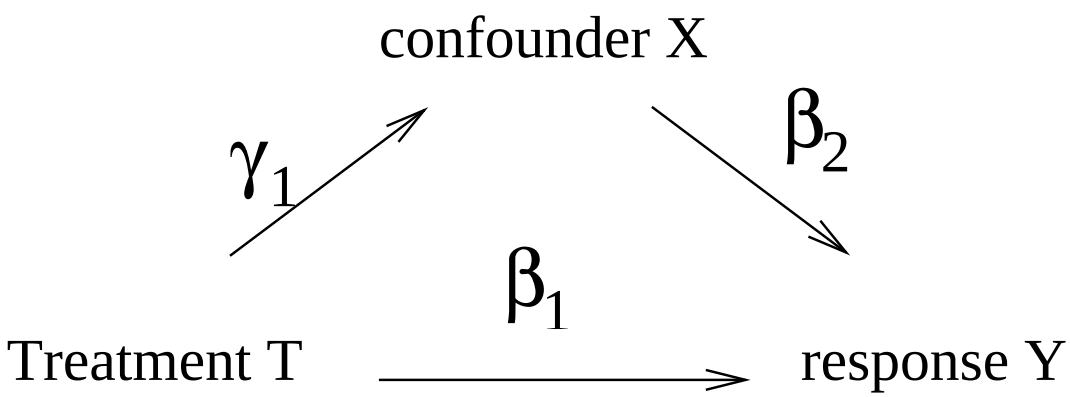
Algebraic explanation

Suppose a “true” model has a treatment T and a confounder x for outcome y.

$$y_i = \beta_0 + \beta_1 T_i + \beta_2 x_i + e_i$$

If x is related to the treatment, we could write

$$x_i = \gamma_0 + \gamma_1 T_i + \nu_i$$



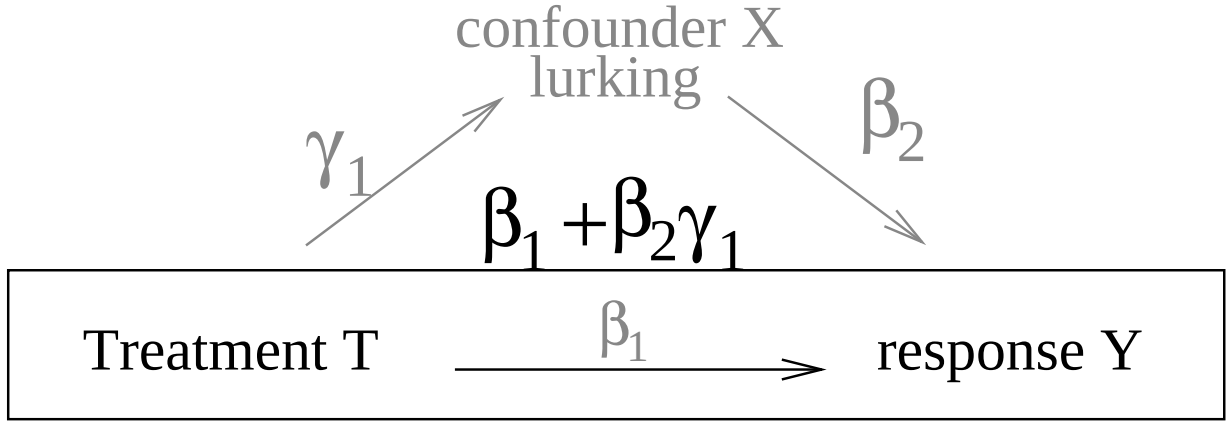
Algebra (cont.) for misleading estimation

If we ignore the confounder x, we would fit the model

$$y_i = \beta_0^* + \beta_1^* T_i + e_i^*$$

The correct model then becomes

$$\begin{aligned} y_i &= \beta_0 + \beta_1 T_i + \beta_2 x_i + e_i \\ &= \beta_0 + \beta_1 T_i + \beta_2(\gamma_0 + \gamma_1 T_i + \nu_i) + e_i \\ &= (\beta_0 + \beta_2\gamma_0) + (\beta_1 + \beta_2\gamma_1)T_i + (\beta_2\nu_i + e_i) \end{aligned}$$



Algebra (cont.)

From before,

$$y_i = (\beta_0 + \beta_2\gamma_0) + (\beta_1 + \beta_2\gamma_1)T_i + (\beta_2\nu_i + e_i)$$

- true treatment effect: β_1.
- estimated effect if x is omitted: $\beta_1^* = \beta_1 + \beta_2\gamma_1$.

Estimation without x is correct only if $\beta_2\gamma_1 = 0$, which happens if

- $\beta_2 = 0$, i.e. x does not predict y, or if
- $\gamma_1 = 0$, i.e. x does not predict T

Causal inference can mislead when there are lurking variables.

FEV example: age as a confounder

```
fit2   = lm(fev~age+smoke, data=fev)
fit1   = lm(fev~    smoke, data=fev)
fitage = lm(    age~smoke, data=fev)

> coef(fit2)       # true beta0, beta2, beta1
(Intercept)          age   smoketrue
  0.3673730   0.2306046  -0.2089949
> coef(fitage)     # gamma0, gamma1
(Intercept)   smoketrue
   9.534805    3.988272
```

Beneficial estimated smoking effect when age is ignored/lurking:

```
> coef(fit1)       # wrong beta0star, beta1star
(Intercept)   smoketrue
  2.5661426   0.7107189

# check that beta0star = beta0 + beta2*gamma0
>  0.3673730 + 0.2306046*9.534805
[1] 2.566143

# check that wrong beta1star = beta1 + beta2*gamma1
> -0.2089949 + 0.2306046*3.988272
[1] 0.710719
```

Factors that affect causal inference

Imbalance and lack of complete overlap can make causal inference difficult for comparing between two treatments.

Imbalance: when treatment groups differ with respect to an important covariate

Lack of complete overlap: when some combination of treatment level and covariate level is lacking (no observations)

Imbalance

- Baby food example: 842 bottle-fed, 274 breastfed. Not same numbers, but no issue.
- Imagine these proportions of girls/boys per food type:

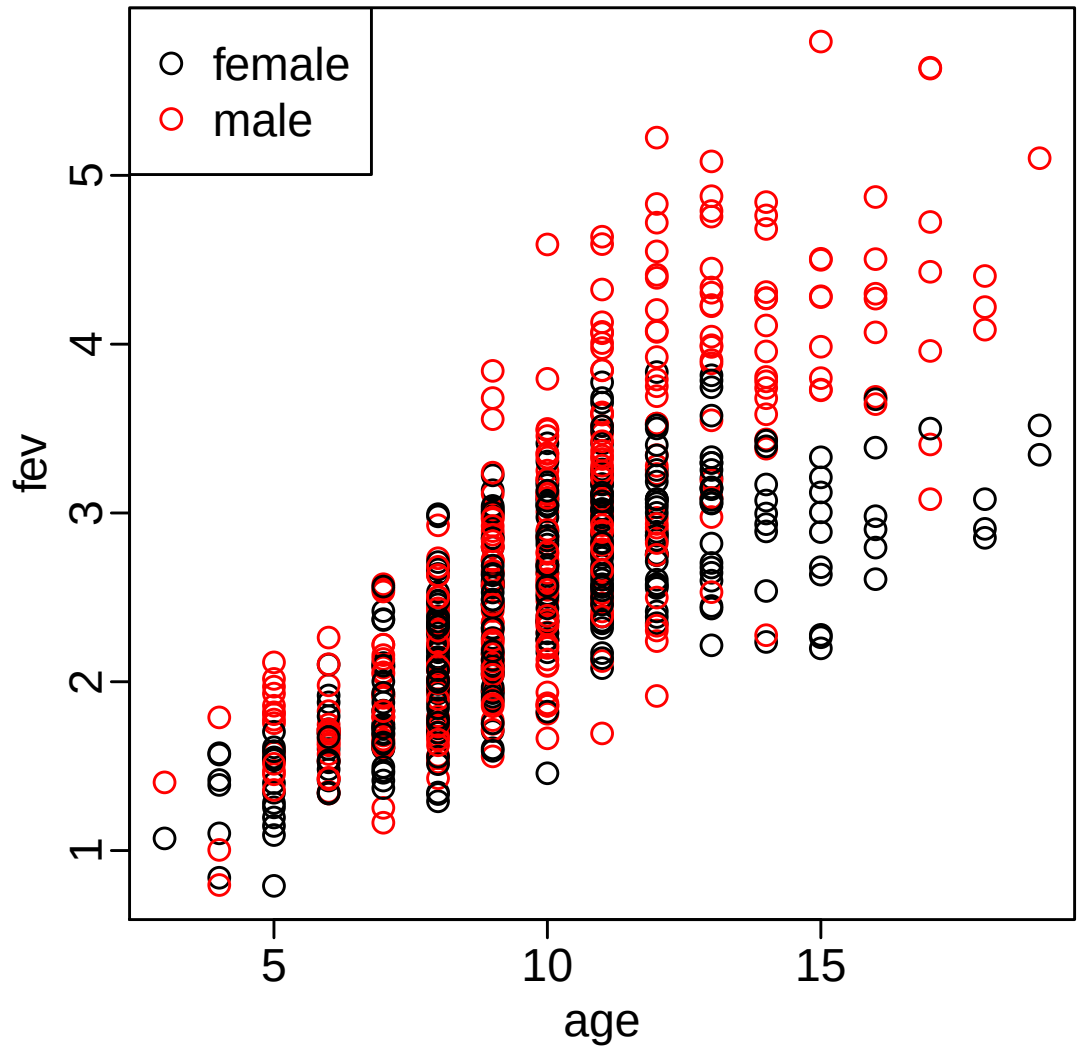
food	boy	girl
bottle-fed	0.80	0.20
breast-fed	0.35	0.65

  There would be a serious difference between groups (bottle/breast-fed) in an important covariate (gender). Gender would be a confounding covariate because it can help predict both the 'treatment' (food) and the outcome (disease).
- We prefer balanced experiments: when all treatment groups are similar with respect to important covariates.
- Randomization is the only way go get balanced groups with respect to *all* covariates, even those we don't suspect!

Imbalance in the FEV example

To explain fev, sex seems to matter, especially among older individuals:

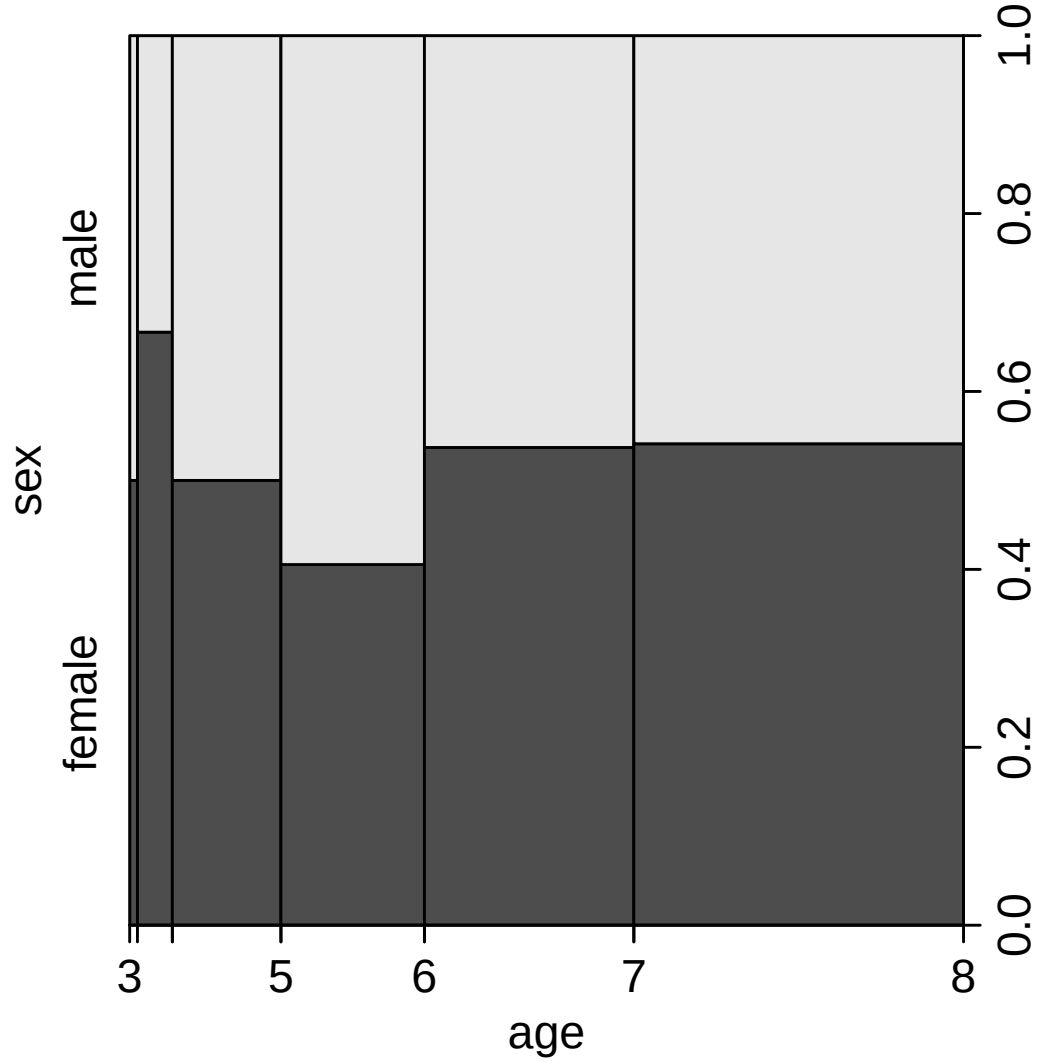
```
plot(fev~age, col=sex, data=fev)
legend("topleft", pch=1, col=1:2, levels(dat$sex))
```



Imbalance in the FEV example

Slight gender imbalance among age categories (for ages < 9)

```
par(mar=c(3,3,1,2), mgp=c(1.4,.4,0), tck=-.02)
plot(sex~age, data=subset(fev, age<9))
```




```
with(subset(fev, age<9),
     table(age, sex)   )
```

age	sex female	male
3	1	1
4	6	3
5	14	14
6	15	22
7	29	25
8	46	39

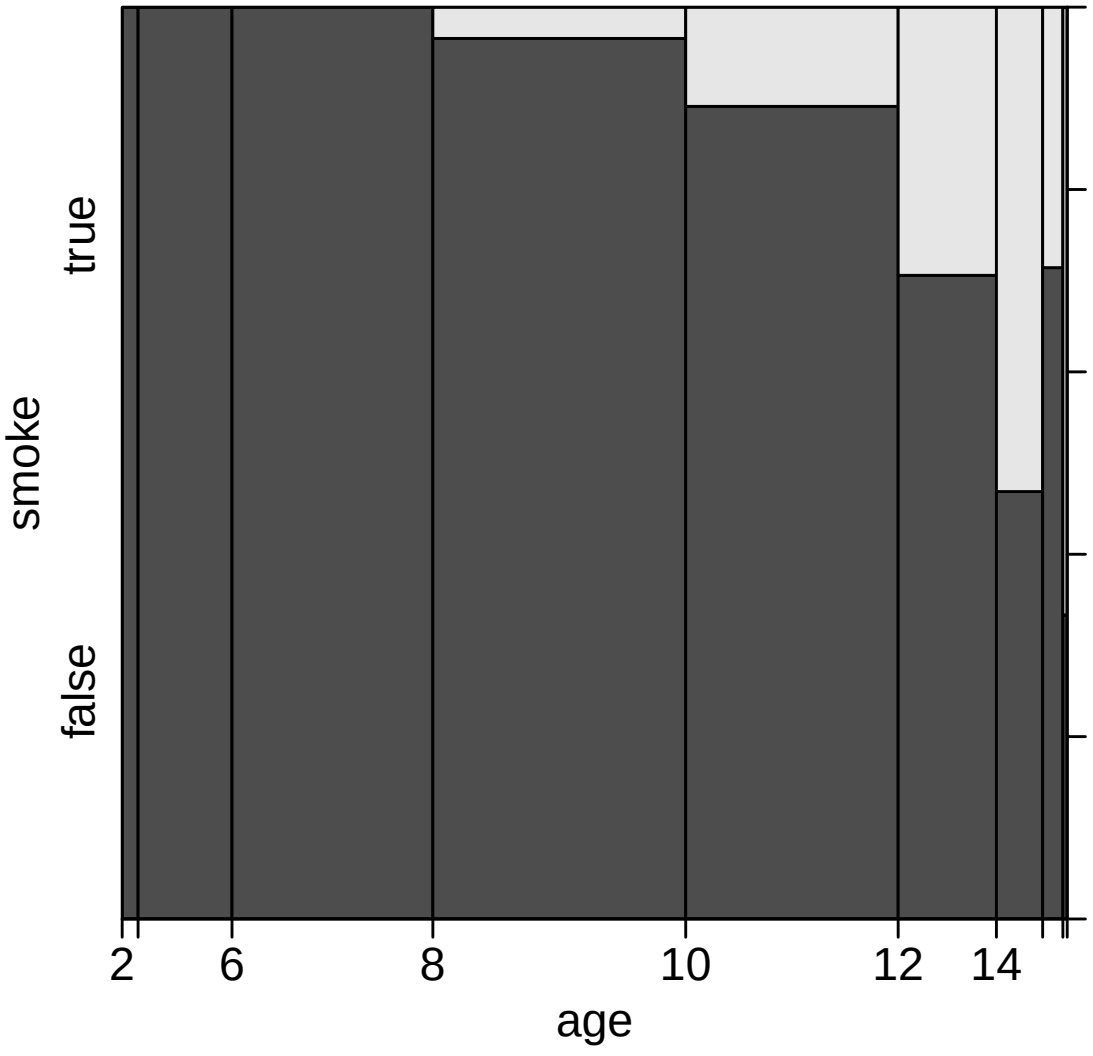
Imbalance

- For imbalanced samples, simple comparisons of sample means between groups are not good estimates of treatment effects.
- After the experiment: a model adjustment is one way to better estimate a treatment effect, where we add the covariate to the model.
- Before the experiment: matching is another strategy to overcome (avoid) imbalance.

  Ex: If gender is thought to affect respiratory disease risk, match each breast-fed baby with a bottle-fed baby of the same gender. Ex: blocking (later).

Lack of complete overlap in the FEV data

Lack of “smoking” children below age 9:




```
> plot(smoke~age, data=fev)
> with(fev,
       table(age, smoke))

    smoke
age  false true
  3      2    0
  4      9    0
  5     28    0
  6     37    0
  7     54    0
  8     85    0
  9     93    1
 10     76    5
 11     81    9
 12     50    7
 13     30   13
 ...         ...
```

Lack of complete overlap

Lack of complete overlap is when there are no observations at some combination(s) of treatment levels / covariate levels.

- For lack of complete overlap, there is no data available for some comparisons.
- This requires extrapolation using a model to make comparisons.
- This is a more serious problem than imbalance.

Fundamental problem of causal inference

- The fundamental problem of causal inference is that at most one of y_i^0 and y_i^1 can be observed.
- Causal inference is predictive inference in a potential-outcome framework.
- Estimation of causal effects requires some combination of:
  - close substitutes for potential outcomes;
  - randomization;
  - or statistical adjustment.

Close substitutes

Several ways to attempt to use close substitutes:

- the same unit can be measured for all treatments (*but are the effects the same*?)
  Ex: treatment 1 in the first week, tmt 2 in week 2, etc. But at least randomize the order.
- a unit can be subdivided into groups and subjected to different treatments (*but do the parts behave identically?*)
  Ex: plots, farms, etc.
- a pre-treatment measurement can be used as a proxy for the control measurement, (*but would the measure have stayed the same under the control?*)
- units are matched with pairs that are very similar on covariates, as in twin studies (*but are the two units truly interchangeable?*)

Blocking

Example: compare alkaloid content in a certain plant species, among cultivated and wild populations.
Design: choose locations where we can find cultivated populations as well as wild populations at the same site.
Imbalance: if the proportion of wild populations varies between sites.

Example: compare milk yield between 2 feeding treatments.
Design: sample farms at random among a certain population of farms. In each farm, allocate a randomly chosen half of the herd to each treatment.

Randomization

A different approach is to use randomization to assign units to treatment groups.

Basic idea: with sufficiently large treatment and control groups, random assignment will result in groups that are essentially very similar in every potential confounder, known or not, and that differences between groups are a measure of an average treatment effect.

Randomization

- Cleanest scenario:
  - individuals are sampled at random from a population;
  - sampled individuals are allocated at random to treatment groups.

  We cannot measure individual level causal effects $y_i^1 - y_i^0$. Instead, we can estimate the

  $$\text{average treatment effect} = \text{mean}(y_i^1 - y_i^0)$$

  Each group acts as a counterfactual for the others: a what if the other treatment were given group.
- More typically:
  - random allocation of treatments to units
  - that are not sampled at random from some population.

  more assumptions are needed to generalize to larger groups.

Completely randomized design (CRD)

Suppose 3 rows and 5 columns of plots in a field. How to assign each plot to one of 3 fertilizers?
Randomly assign fertilizers to the $3n = 15$ plots. Example:

```
> rep(1:3, each=5)
 [1] 1 1 1 1 1 2 2 2 2 2 3 3 3 3 3
> sample(rep(1:3, each=5))
 [1] 1 3 3 2 2 3 1 1 2 2 1 2 3 1 3
```

1	3	3	2	2
3	1	1	2	2
1	2	3	1	3

Given this design, what is the most appropriate analysis to compare the 3 fertilizers?

Randomized complete block design (RCBD)

Suppose an east-west slope, or east-west gradient in some factor that is known to affect the response. Then we would 'block' together all plots on the farthest east side, etc. Example:

```
> sample(1:3)
[1] 1 3 2
> replicate(5, sample(1:3))
     [,1] [,2] [,3] [,4] [,5]
[1,]    1    3    2    3    3
[2,]    2    1    3    1    1
[3,]    3    2    1    2    2
```

Goal: control known sources of variability and reduce error within blocks. Suitable when units are heterogeneous and can be grouped into homogeneous units within blocks.

Randomized complete block design (RCBD)

Given this design, what is the most appropriate analysis to compare the 3 fertilizers?

```
lm(response ~ block + fertilizer)
```

where ‘block’ is the column ID, categorical. Here, this is a two-way ANOVA with the blocking factor as a fixed effect.

The block effect would be better considered as a random effect, more on this later.

Statistical adjustment

- When groups are not similar, regression of other statistical models can be used to estimate what the outcome might have been under a different treatment.
- The sample could be subdivided into subsets where group allocation mimics a randomized experiment.

Example: alkaloid content in wild and cultivated populations at different sites. Think what variables might be lurking, and incorporate them into the analysis.

Statistical adjustment

Example: due to space constraints, an experiment must be split between 2 greenhouses.
Should we assign the treated plants to the first greenhouse, and the control plants to the second greenhouse? Randomize (how?) Block (how?) or use statistical adjustment after the fact?

- Imbalance:
- Lack of complete overlap:
- Complete lack of overlap:

Bottom line

Randomized experiments are better able to eliminate lurking variables than observational experiments.

All forms of causal inference depend on more assumptions for validity than does predictive inference.