Multivariate Relationships and Stratification

Ryan Miller

Grinnell College
Statistics

Introduction

Many research questions can be distilled to assessing a *bivariate* relationship:

> *How does the explanatory variable X relate with the response variable Y*

- Unfortunately, the relationship between X and Y can be influenced by other variables
  - It's even possible for a third variable to be so influential that it completely reverses the direction between X and Y, a phenomenon known as **Simpson's Paradox**

Confounding

The relationship between variables X and Y is said to be **confounded** by a third variable, Z, if the variable Z is associated with *both* X and Y

Confounding

The relationship between variables X and Y is said to be **confounded** by a third variable, Z, if the variable Z is associated with *both* X and Y

- In the death penalty sentencing data, victim's race was a *confounding variable* in the relationship between offender's race and death penalty sentence
  - White offenders were mostly involved in cases with white victims (ie: X and Z are associated)
  - Cases with a White victim were much more likely to result in a death penalty sentence than cases with a Black victim (ie: Z and Y are associated)

Conditional Effects and Stratification

If we identify a *confounding variable* in our analysis, we'll want to *control* for it:

1. **Stratification** controls for a categorical confound by describing the association between the explanatory and response variables separately for each group created by the confounding variable.
2. **Multivariable regression**, our next topic, provides another way to control for confounding variables.

Conditional Effects and Stratification

- In the death penalty sentencing example, we might report separate odds ratios for cases involving white victims and cases involving black victims
  - These odds ratios *condition* on a value of the confounding variable, for example if we *condition on the victim being white* the odds ratio of a death penalty verdict for black offenders relative to white offenders is calculated: $\text{OR}_{\text{white vic}} = \frac{37/41}{46/144} = 2.825$

Table 1: 'White Victim'

	death	not
black	37	41
white	46	144

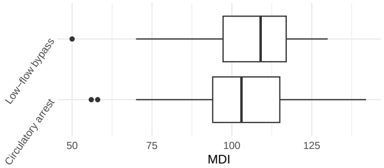
Table 2: 'Black Victim'

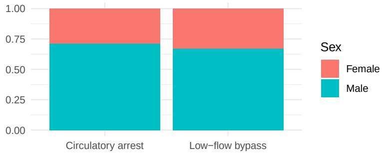
	death	not
black	1	101
white	0	8

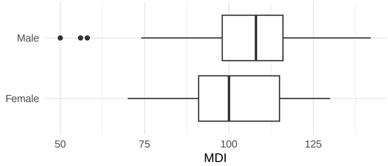
Grinnell College
Statistics

Is it confounding?

The "infant heart" data come from a Harvard Medical School study comparing two types of surgical treatments for infants born with congenital heart defects. Does *sex* confound the relationship between *treatment* and *mental development score (MDI)*:







Is it confounding?

- The variables "MDI" and "sex" *do* appear to be associated (ie: Y and Z are associated)
    - But "sex" and "treatment" *do not* appear to be associated (ie: Y and X are not associated)

Is it confounding?

- The variables “MDI” and “sex” *do* appear to be associated (ie: Y and Z are associated)
  - But “sex” and “treatment” *do not* appear to be associated (ie: Y and X are not associated)
- Thus, the definition of confounding is not met, as our third variable, “sex”, is not associated with the explanatory variable, “treatment”
  - So, the observed difference in MDI scores between the low-flow and the circulatory groups cannot be explained by the variable “sex”
  - Any impact of “sex” on “MDI” occurs evenly across the two groups

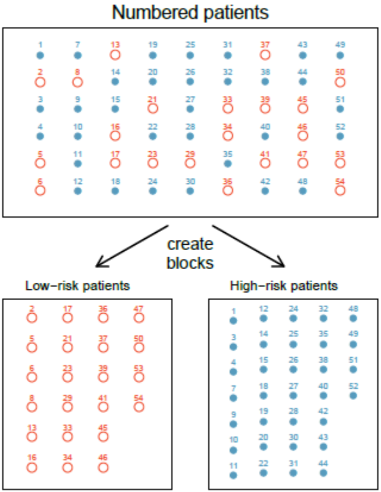
Randomized experiments

- It is possible to *force* an explanatory variable to have *no association* with a third variable using **random assignment**
  - In the infant heart study, researchers randomly assigned a surgery to each infant
  - Thus, male infants were equally likely to receive either type of surgery, which is why we did not see a significant relationship between the variables "sex" and "treatment"

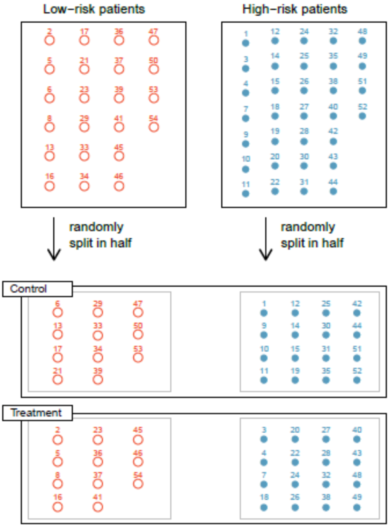
Randomized experiments

- It is possible to *force* an explanatory variable to have *no association* with a third variable using **random assignment**
  - In the infant heart study, researchers randomly assigned a surgery to each infant
  - Thus, male infants were equally likely to receive either type of surgery, which is why we did not see a significant relationship between the variables "sex" and "treatment"
- Random assignment doesn't *guarantee* that groups will be *exactly equal*, but it does ensure that they'll be balanced across all confounding variables *on average*
  - If a variable is very strongly associated with the outcome, researchers can ensure it is *exactly balanced* using **blocking**

Blocking - first create blocks



Blocking - then randomly assign within blocks



Limitations

- Random assignment is not always feasible
  - Some explanatory variables, such as demographic or genetic factors, cannot be randomly assigned
  - Randomized experiments are also much more costly than other types of studies
- Stratification is only viable when the data include a small number of categorical confounds
  - As few as three binary categorical confounds would lead to $2^3 = 8$ different sets of conditional descriptive statistics
- Our next topic, **multivariable regression**, offers a more flexible method of controlling for multiple confounding variables