STAT718/BIOL703: Genomic Data Science
Confounding Effect

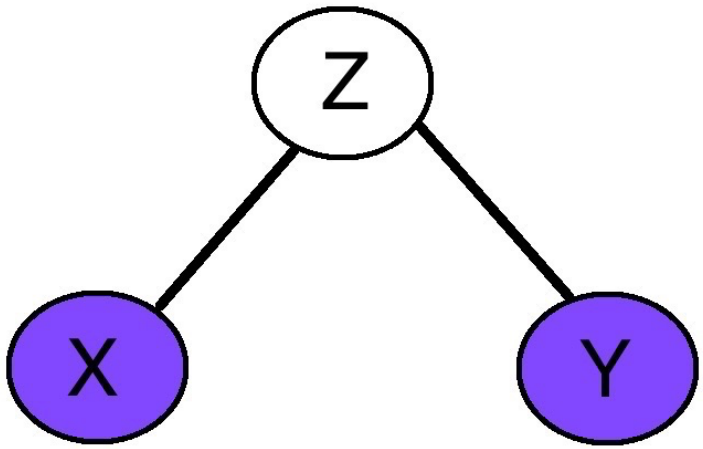
Dr. Yen-Yi Ho (hoyen@stat.sc.edu)

Confounding: Ice cream consumption and drowning death

A researcher looked into a public database from a southern city, and found an association between ice-cream consumption and number of drowning deaths for a given period. He concluded that ice cream could cause drowning.

Do you agree? Why do you think the researcher found such association?

Confounding



- Associated with X
- Independently associated with Y
- Not in the causal pathway between X and Y
- ex: ice cream consumption and drowning death
- ex: Chopsticks gene

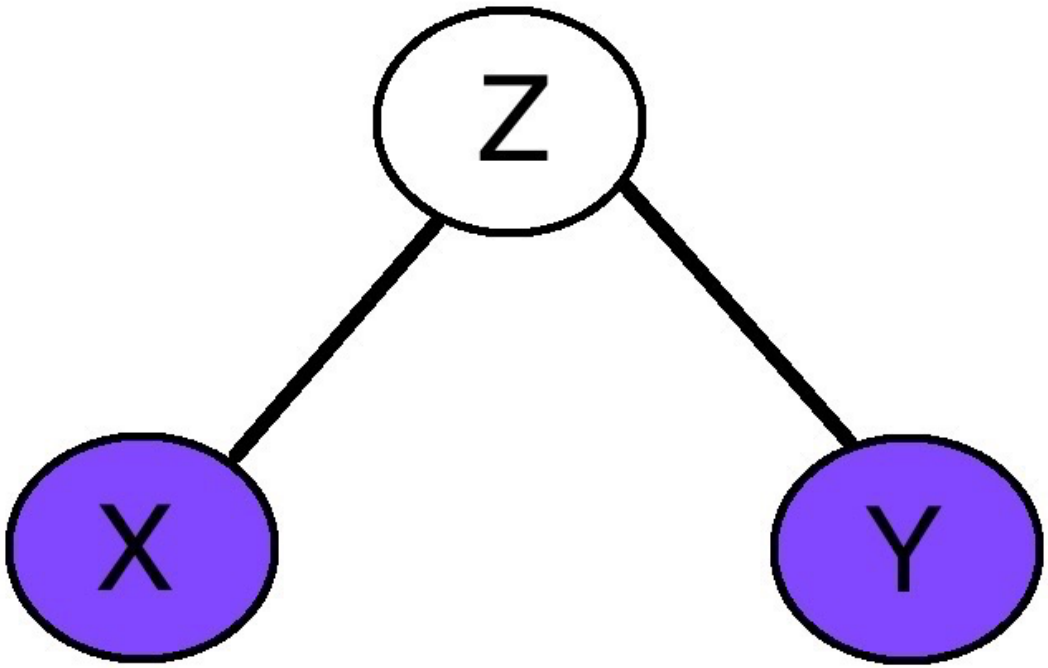
Confounding

Once upon a time, an ethnogeneticist decided to figure out why some people eat with chopsticks and others do not. His experiment was simple. He rounded up several hundred students from a local university, asked them how often they used chopsticks, then collected buccal DNA samples and mapped them for a series of anonymous and candidate genes.

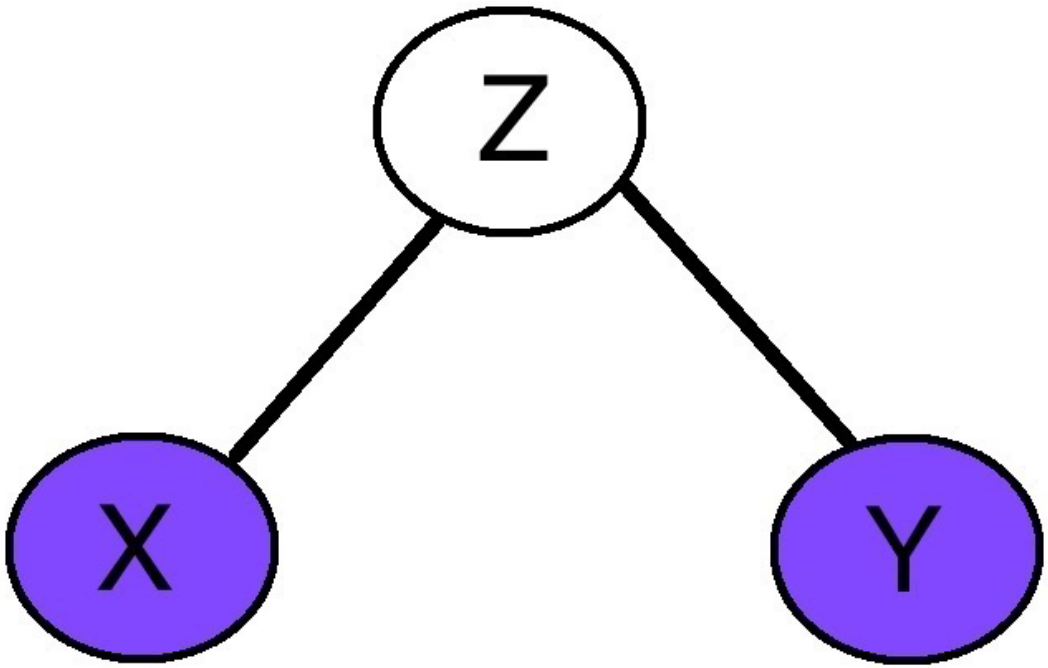
Confounding

The results were astounding. One of the markers, located right in the middle of a region previously linked to several behavioral traits, showed a huge correlation to chopstick use, enough to account for nearly half of the observed variance. When the experiment was repeated with students from a different university, precisely the same marker lit up. Eureka! The delighted scientist popped a bottle of champagne and quickly submitted an article to Molecular Psychiatry heralding the discovery of the 'successful-use-of-selected-hand-instruments gene' (SUSHI).

Confounding



Confounding



To control for potential confounding effect, we need to stratified by racial groups and perform the analysis within each group.

More Example

- Wikipedia's entry on Simpson's paradox gives an example comparing two player's batting averages

	First Half	Second Half	Whole Season
Player 1	4/10 (.40)	25/100 (.25)	29/110 (.26)
Plater 2	35/100 (.35)	2/10 (.20)	37/110 (.34)

- Player 1 has a better batting average than Player 2 in both the first and second half of the season, yet has a worse batting average overall
- Consider the number of at-bats

Berkeley admissions data

- The Berkeley admissions data is a well known data set regarding Simpsons paradox

```
?UCBAdmissions
data(UCBAdmissions)
     apply(UCBAdmissions, c(1, 2), sum)
          Gender
Admit      Male Female
  Admitted 1198    557
  Rejected 1493   1278
           .445   .304 <- Acceptance rate
```

Acceptance rate by department

```
> apply(UCBAdmissions, 3,
        function(x) c(x[1] / sum(x[1 : 2]),
                      x[3] / sum(x[3 : 4])
                      )
        )
Dept   M    F
   A 0.62 0.82
   B 0.63 0.68
   C 0.37 0.34
   D 0.33 0.35
   E 0.28 0.24
   F 0.06 0.07
```

Discussion

- Mathematically, Simpson's paradox is not paradoxical

$$\begin{aligned} a/b &< c/d \\ e/f &< g/h \\ (a+e)/(b+f) &> (c+g)/(d+h) \end{aligned}$$

- More statistically, it says that the apparent relationship between two variables can change in the light or absence of a third

Confounding

- Variables that are correlated with both the explanatory and response variables can distort the estimated effect
- One strategy to adjust for confounding variables is to **stratify** by the confounder and then combine the strata-specific estimates
    - Requires appropriately weighting the strata-specific estimates
- Unnecessary stratification reduces precision

Hierarchy for Scientific Evidence in Medical Studies

On 15 July 1911, 65-year-old Mrs. Jane Decker was struck by lightning while in her house. She had been deaf since birth, but after being struck, she recovered her hearing.

Is this compelling evidence that lightning is a cure for deafness?

Hierarchy for Scientific Evidence in Medical Studies

- Observational study
  - Anecdotal evidence: a short story or an example of an interesting event.
  - Retrospective case-control studies (case-control)
  - Prospective cohort studies
- Experiment
  - Randomized clinical trials

Hierarchy for Scientific Evidence in Medical Studies

Smoking $\rightarrow$ lung cancer

- **Anecdotal evidence** example: My uncle smoke and he got lung cancer.
- **Retrospective case-control study** example: 100 lung cancer patients and 100 normal control were enrolled; then smoking history for each participants was examined.
- **Prospective cohort study** example: 100 smokers and 100 nonsmoker were enrolled, then were followed up to determine whether they developed lung cancer.
- **Randomized clinical trial** example: Participants were randomized to receive low-nicotine cigarette or regular cigarette; then were followed-up to determine whether they developed lung cancer.