Confounding and Adjusting

Comparisons need to be *fair*.

In observational studies the comparisons are usually not fair.

Other variables are *confounded* with the variable of interest.

This is often dealt with by trying to *adjust* the analysis for the confounding variables.

You can only do this if you know what confounders are important.

AND you measure them.

Upstate New York data on Prostate Cancer deaths in 1994.

White men: 1359 deaths

Black men: 121 deaths

Comparison?

Can't be done, yet.

Need to convert to *rates*.

Compare numbers of deaths to number of men in each race.

	Race	
	White	Black
# Cases	1359	121
# Men	4,738,246	418,992

Convert to rates per hundred thousand men:

White cases per 100,000 men:

$$\frac{1359}{4,738,246} \times 100,000 = 28.7$$

Black cases per 100,000 men:

$$\frac{121}{418,992} \times 100,000 = 28.9$$

No difference?

Hypothesis test for equality of two proportions:

$$z = \frac{\widehat{p}_W - \widehat{p}_B}{\sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{n_W} + \frac{1}{n_B}\right)}}$$

Get $z = -0.07$ and two-sided $P = 0.94$. No evidence of difference.

Now look at younger men separately from older men.

Age is *known* to be confounded with these variables (race and prostate cancer).

Older men are more likely to get prostate cancer.

Black men are likely to be younger.

Split into under / over 65 years of age.

	Young		Old	
	White	Black	White	Black
Cases	76	18	1282	102
Men	4,177,889	396,917	560,357	22,075
Rate	1.8	4.5	228.8	462.1

Simpson's paradox:

1. The number of young black men is nearly 10% of the number of young white men.

2. But old black men less than 5% of number of old white men.

3. In both subgroups the rates for black men are far higher than for white men.

4. *Adjusted for age* or *controlling for age* prostate cancer rates are far higher for black than for white men.

5. Overall rates comparable because black men are younger and younger men suffer from less prostate cancer.

6. Not a real paradox – just a fact about how arithmetic can work out.

Sex bias in graduate school admissions

Bickel et al. *Science*, 1975

	Men	Women
# Accepted	3738	1494
# Refused	4704	2827
Acceptance Rate	44.3%	34.6%

Success rate is far higher for men than women.

Bias in the admission process?

Which department?

Six largest majors:

	Men		Women	
Major	Appl	Admit %	Appl	Admit %
A	825	62	108	82
B	560	63	25	68
C	325	37	593	34
D	417	33	375	35
E	191	28	393	24
F	373	6	341	7

Now test independence between Admission and Sex in each department.

Department A: Observed Table

	Men	Women
# Accepted	511	89
# Refused	314	19

Expected Table:

	Men	Women
# Accepted	530.5	69.5
# Refused	294.5	37.5

$$X^2 = \frac{(511 - 530.5)^2}{530.5} + \frac{(89 - 69.5)^2}{69.5} + \frac{(314 - 294.5)^2}{294.5} + \frac{(19 - 37.5)^2}{37.5} = 16.6$$

$$P = 0.000046$$

Conclusion: random sampling cannot easily explain the different success rates for men and women.

Discussion of tactics:

Some majors harder to get into than others.

Women tend to pick hard majors to get into.

So Department Choice is confounded with Sex.

Adjust for Department choice by making comparisons within departments.

Simpson's paradox: conclusion of the adjusted analysis can be quite different than that of aggregate analysis.

All examples of a general tactic for analyzing observational studies:

Use *disaggregation* or *cross-classification* or *regression* to adjust the effect of X on Y for the effect of some known confounding variable Z.

Association between variables

A major question of research interest: are two variables related?

NOTE: if yes we need to go on and say HOW they are related.

So far:

One binary variable (like sex) with one continuous variable: two sample or paired comparisons t-test.

Example: is breeding method related to plant height?

Answered by comparing means between two breeding methods.

Two binary variables: two sample test for equal proportions.

Example: mantis colour (one variable), survival (other variable).

Two categorical variables: test of independence in a contingency table.

Example: streams – flow rate related to bed type?

Next case: two continuous variables – *regression*.

Inference in Regression

Chapter 23: my focus – how to do it.

Sample of 534 from Current Population Survey.

Several variables: education, age, sex, experience, wages, race, union status , etc.

Which of these are related to Wages?

First do simple problem: regression of education on age.

Use JMP in class to do regression.

Output shows

```
EDUC = 14.251125 - 0.0334588*AGE
```

Slope and intercept are *estimates* of population quantities.

Have standard errors.

$b = -0.0334588$, $SE_b = 0.00956$

Can test hypothesis that two variables are unrelated in population:

$$H_o : \text{Population slope}=0$$

Test statistic is

$$t = \frac{b - 0}{SE_b} = -3.50$$

Get P value from t distribution with 534-2=532 degrees of freedom.

Find

$$P = 0.0005$$

The variables are related and older people have less education.

Confidence intervals are also possible:

$$b \pm t^* SE_b$$

For our data with 95% confidence we want $t^* = 1.644$ and our interval is

$$-0.0335 \pm 1.644 \times 0.00956.$$

We are 95% confident that the slope of the *population* regression line is in this range.

Warning: There are assumptions behind the method.

Is the relationship in the population a straight line?

Is the variability in y for a fixed x constant?

Are the data points sampled independently from a population?

First two questions studied with data using plots of residuals.

Residual plots suggest problem with constant variability.

Logic suggests problem with straight line (young people with 18 years of educ?).

Adjustment by regression

So far: confounding variable nominal or categorical.

If confounding variable continuous we use *regression* to adjust.

Do example in JMP: effect of sex on wage. Adjust for experience.